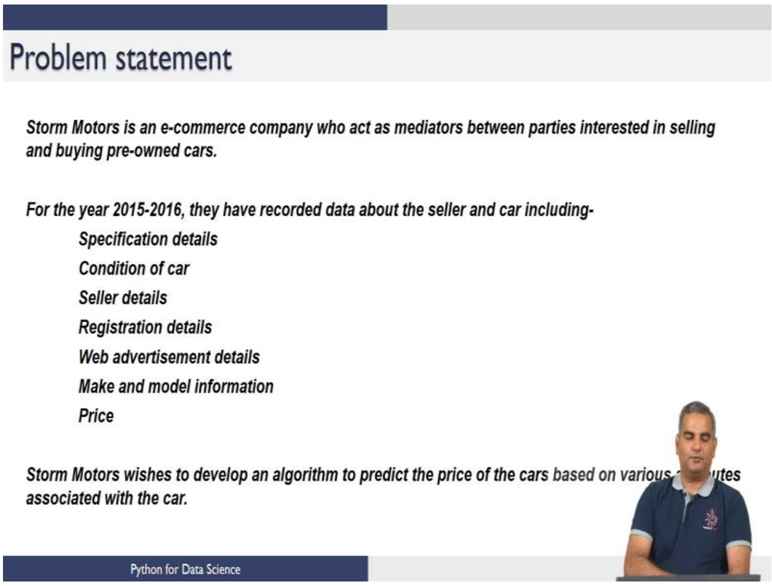
Python for Data Science
Prof. Prathap Haridoss
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 42
Introduction to Regression Case Study

So, the previous case study that we introduced was a classification case study and as a second case study we are going to Introduce a Regression or a function approximation case study. It basically follows the same format as the classification case study. So, I am going to quickly go through this case study introduction so that you can see how this case study is solved in Python.

So, the case study is on predicting the price of pre-owned cars. So, right at the beginning we can see that here there is a slight difference between what we saw in the last case study and this. In the last case study we simply wanted to classify individuals into two possible categories people who are making less than 50,000 and people who are making more than 50,000 whereas, here what we are looking to do is we are actually trying to see if we can predict the price of pre-owned cars. So, there is just no category, it is a value for a pre-owned car and how do we predict the price of pre-owned cars.

(Refer Slide Time: 01:31)



Problem statement

Storm Motors is an e-commerce company who act as mediators between parties interested in selling and buying pre-owned cars.

For the year 2015-2016, they have recorded data about the seller and car including-

- *Specification details*
- *Condition of car*
- *Seller details*
- *Registration details*
- *Web advertisement details*
- *Make and model information*
- *Price*

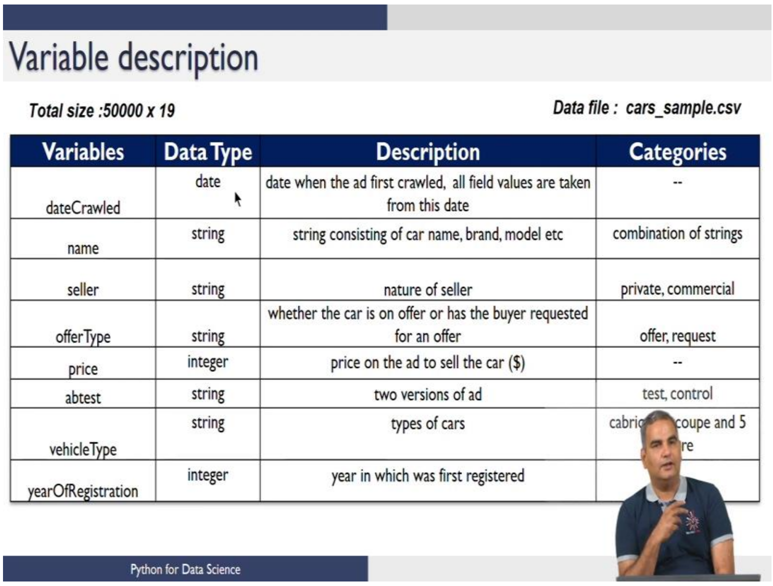
Storm Motors wishes to develop an algorithm to predict the price of the cars based on various attributes associated with the car.

So, the problem statement is a Storm Motors is an e-commerce company, then they act as mediators between parties interested in selling and buying pre-owned cars and they have a lots of data based on sales that have happened through them or otherwise and what the interest is in is to make a sale quickly. So, if the price is appropriate and that is something that you can satisfy both the seller and the buyer, then you will have your cars moving fast.

So, what Storm Motors wants to do is they want to develop an algorithm, so that they can predict the price of the cars based on various attributes associated with the car. So, you might think of using a such a model from a Storm Motors viewpoint in multiple ways. If if a seller is asking for a price you could put the attributes of the car that the seller is selling into this model and then come up with a predicted price, and if it is much above what it is predicting then you can tell the seller that you know you are you are asking for too much and you are not likely to sell this car at this price.

And, similarly a buyer comes and bids for a car at a much lesser price than what has been put up, then you could show the results of the model and tell the buyer, look you are asking for a very cheap price so, the seller is unlikely to give you this car at this price. So, if you got this right then you could optimize this transaction and then have a both the parties being happy and then you have a better business.

(Refer Slide Time: 03:33)



Variable description

Total size :50000 x 19

Data file : cars_sample.csv

Variables	Data Type	Description	Categories
dateCrawled	date	date when the ad first crawled, all field values are taken from this date	--
name	string	string consisting of car name, brand, model etc	combination of strings
seller	string	nature of seller	private, commercial
offerType	string	whether the car is on offer or has the buyer requested for an offer	offer, request
price	integer	price on the ad to sell the car ($)	--
abtest	string	two versions of ad	test, control
vehicleType	string	types of cars	cabric coupe and 5 re
yearOfRegistration	integer	year in which was first registered	

Python for Data Science

And, much like what we saw in the classification problem again a we think about this data in a matrix format. So, Storm Motors has data for about 50,000 cars in their database that have been sold or that have been processed in one way another and there are these 19 variables that are associated with this problem. Clearly, one of these variables is going to be the outcome variable or the price of the car and the other variables are variables that we hope have enough information in them so that we can basically predict the price of the car.

And, much like in the last example if we go through these variables so, there is a variable called dateCrawled and the data type is date. So, this is a variable where a once Storm Motors puts the ad out when did it catch the first eyeball. So, when was the first time this ad was crawled and all field values taken from this data, so, that date is the first date crawl. So, it gives you an idea of when people roughly started looking at a this car.

Now, the name is another variable and this is kind of a composite variable. This is been downloaded from someplace. So, this is a string variable, but this is a string variable that could consist of car name, brand, model etcetera. So, it is kind of a composite string and in this data set you will see that it is not always consistent in terms of the order in which all of the information comes in.

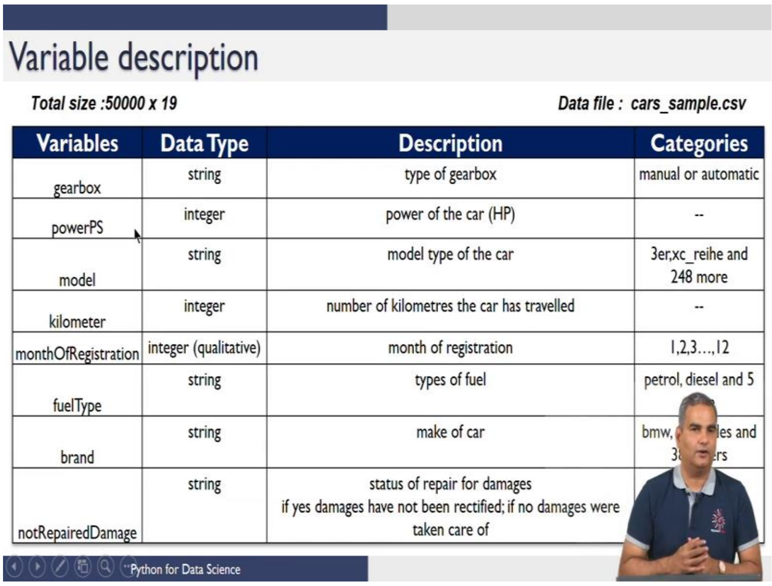
Seller whether it is a private or a commercial seller; offer type is whether the buyer has looked at a particular car and then gave an offer saying this is your price, but I am willing to buy this car for this price or it is a price that the seller has asked for. So, that is the other variable that we have. Price is the price on the ad to sell the car which is the outcome variable that we are interested in and the way these ads are put in there are certain characteristics of these ads and as part of this exercise Storm Motors also had some specific studies that they want to conduct.

So, the ads could be of the test type or the control type and the data type is a string vehicle type is a string whether it is a Cabrios, SUV, Coupe and 5 more different types of vehicle; year of registration year in which the car was first registered which is an integer variable.

Now, you can see that many of these clearly will have an impact on what price you can sell the car right. So, for example, vehicle type is an important aspect one would one would assume and there are more variables that we will see; seller for example, private or commercial could have an impact or might not. So, this is something that we might not know; private sellers might want to hold on and get the best price for the car, sometimes commercial guys might want to move cars out. So, we really do not know how much an impact it will have, but these are all relevant variables.

So, whenever you look at a data sense problem you basically look at whether the variables are relevant to the problem and then how relevant and quantitatively, how much the impact and so on, only the data will tell you. So, that is post the solution you will be able to understand.

(Refer Slide Time: 07:37)



Variable description

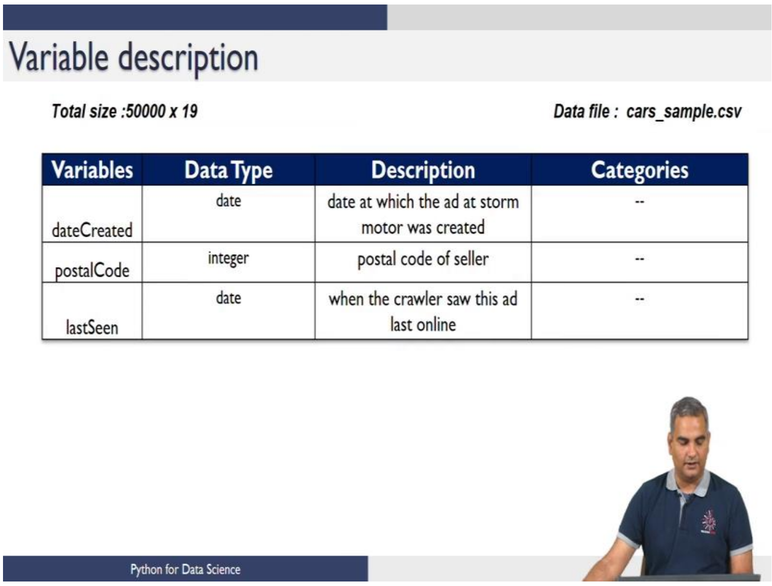
Total size :50000 x 19 *Data file : cars_sample.csv*

Variables	Data Type	Description	Categories
gearbox	string	type of gearbox	manual or automatic
powerPS	integer	power of the car (HP)	--
model	string	model type of the car	3er,xc_reihe and 248 more
kilometer	integer	number of kilometres the car has travelled	--
monthOfRegistration	integer (qualitative)	month of registration	1,2,3…,12
fuelType	string	types of fuel	petrol, diesel and 5 [illegible]
brand	string	make of car	bmw, [illegible]les and 3[illegible]ers
notRepairedDamage	string	status of repair for damages if yes damages have not been rectified; if no damages were taken care of	[illegible]

Now, going on to more variables and there is a variable called gearbox which is string, it can be manual or automatic; powerPS – power of the car and this is an integer and there are multiple values this can take. Model of the car model type of the car for example, if it is a Hyundai is it an i10, i20 and so on; kilometre – number of the number of kilometers the car has travelled which is an integer variable. Month of registration, fuelType whether it is a patrol car, diesel car and 5 more. Brand – what type of car is it, is it a BMW, Mercedes and many others.

Now, we also have another variable which might be important from a price of car view point: notRepairedDamage is a variable name. So, this is a string if it is yes, then basically what it says is there has been a damage and it is not been repaired. So, are not rectified and no means there has been a damage, but the car has been repaired and rectified. So, you can again see how this variable might have an impact in terms of the final price.

(Refer Slide Time: 09:03)



Variable description

Total size :50000 x 19

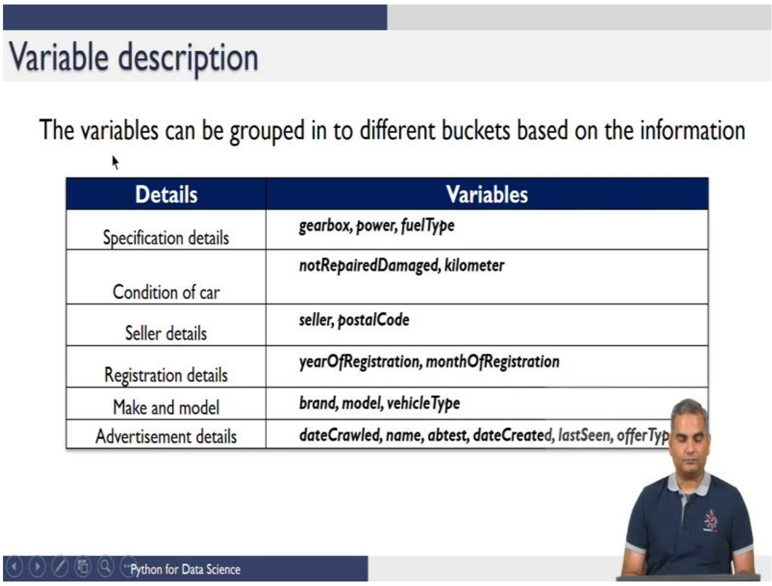
Data file : cars_sample.csv

Variables	Data Type	Description	Categories
dateCreated	date	date at which the ad at storm motor was created	--
postalCode	integer	postal code of seller	--
lastSeen	date	when the crawler saw this ad last online	--

Python for Data Science

So, dateCreated another variable of data type date data to with that Storm Motors was created; postalCode – postal code of the seller and lastSeen – when the crawler saw this ad last online. So, what has been the most recent activity in terms of interest in terms of this car.

(Refer Slide Time: 09:23)



Variable description

The variables can be grouped in to different buckets based on the information

Details	Variables
Specification details	*gearbox, power, fuelType*
Condition of car	*notRepairedDamaged, kilometer*
Seller details	*seller, postalCode*
Registration details	*yearOfRegistration, monthOfRegistration*
Make and model	*brand, model, vehicleType*
Advertisement details	*dateCrawled, name, abtest, dateCreated, lastSeen, offerTyp*

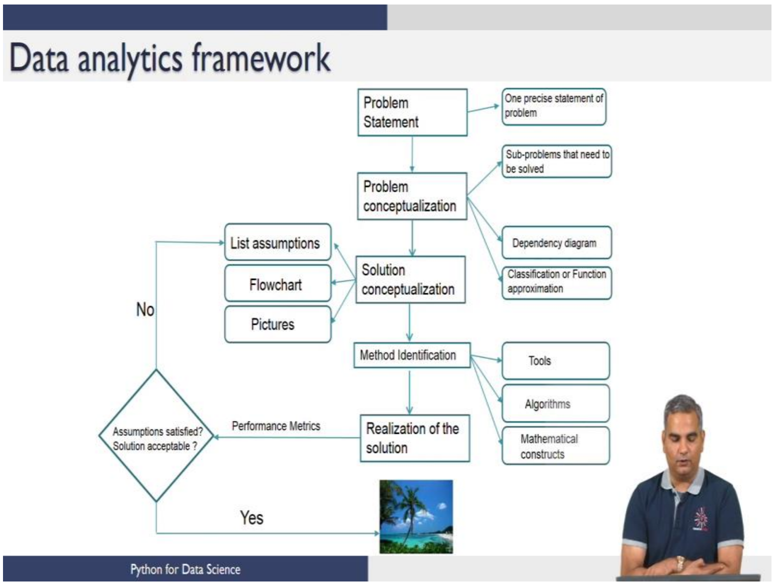
Python for Data Science

Now, you can see in kind of think about this and then say each of these variables could have an impact. For example, the last thing that I talked about if a car is not getting any eyeballs at all, then you know maybe it is it is it is not priced right. So, you have to really think about how to price it, so that you can sell it ok. So, based on this information the variables can be a grouped into different buckets which is something that you can think about.

And, say well, the vehicle specification details such as gearbox, power, fuel type is one group; condition of the car and not repair damage and kilometres another group. Both of them in some sense tell you how the car is likely to be in what condition is it likely to be and so on if it is run a large number of kilometres then you know it is a older car so, condition might not be great and of course, not repair damage directly gives you information about the condition of a car.

Seller details again which part of the geography the seller comes from and so on. So, those can be grouped under seller details. Registration details – year of registration, month of registration of course, the car itself make an model and what is happening based on the advertisements that we have and does that give you any clue in terms of what is likely to the price is another group that you can think of.

(Refer Slide Time: 11:05)



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 ?

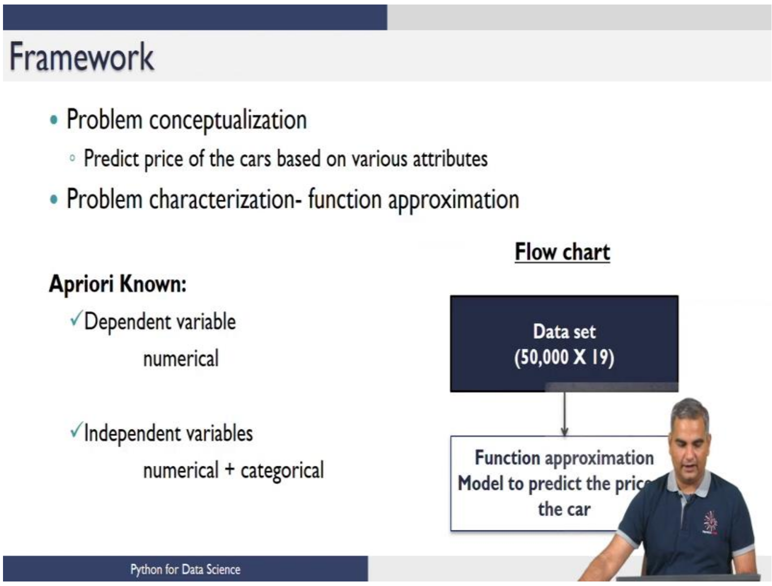
No

Yes

Python for Data Science

Again, I do not want to go through this slide again. We use the same process that we talked about in the classification example. Same notions of problem statement, very clear here. We want to predict a price of a used car and again since it is a very simple straight one problem you do not really need to look at sub-problems and so on. It is a function approximation or a regression problem and much like how we talked about the previous case, again here we are going to run a several different models and then find out whichever model does well and all of this we are going to again do this using Python. So, that is basically what we have.

(Refer Slide Time: 11:47)



Framework

- Problem conceptualization
  - Predict price of the cars based on various attributes
- Problem characterization- function approximation

Apriori Known:

✓Dependent variable

numerical

✓Independent variables

numerical + categorical

Flow chart

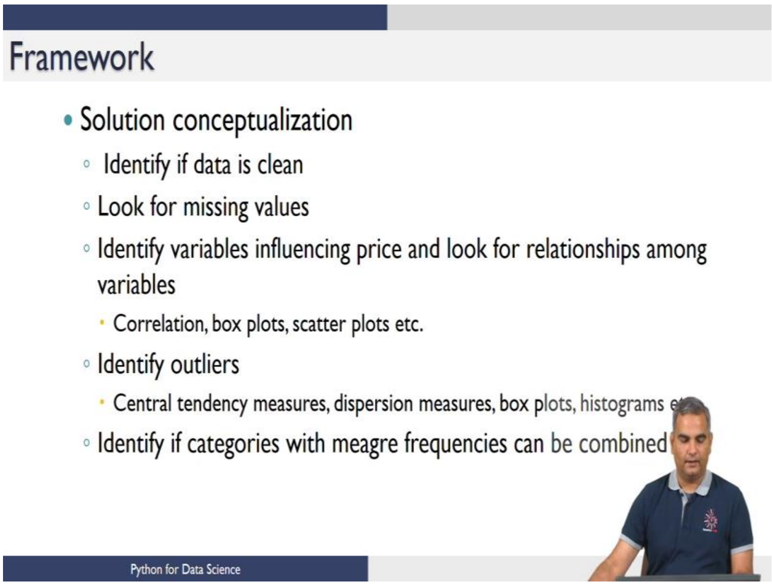
Data set (50,000 X 19)

Function approximation Model to predict the price of the car

Python for Data Science

So, we know that the dependent variable is a numerical variable which is the price of the car that we want to predict and again independent variables have a combination of numerical and categorical variables.

(Refer Slide Time: 12:03)



Framework

- Solution conceptualization
  - Identify if data is clean
  - Look for missing values
  - Identify variables influencing price and look for relationships among variables
    - Correlation, box plots, scatter plots etc.
  - Identify outliers
    - Central tendency measures, dispersion measures, box plots, histograms etc.
  - Identify if categories with meagre frequencies can be combined

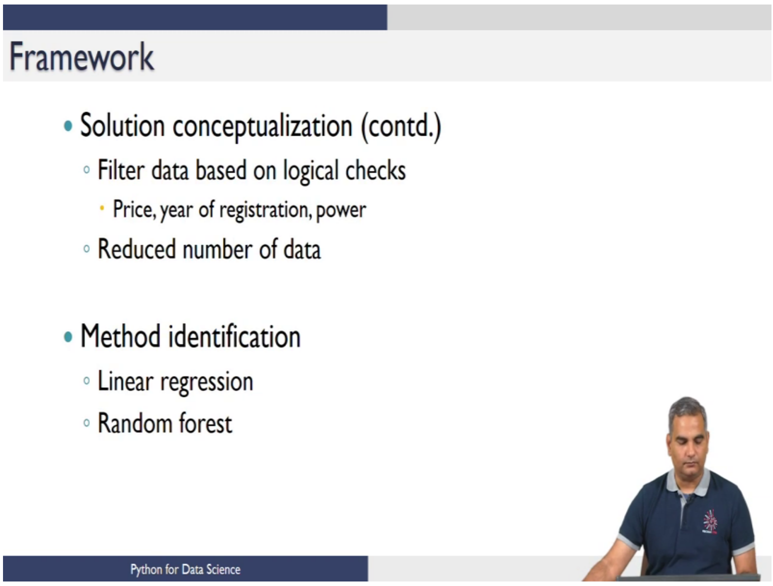
Python for Data Science

Go through a pretty much same kind of ideas that we talked about in the classification problem. First look if the data is clean, look for missing values, look for variables that might influence price. So, this is where you do some form of visualization, descriptive statistics and so on, and also in this case one of the important things is to identify outliers. There are data points which really do not make much sense at all based on whatever logic that we used.

So, for example, if there are a lot of data points which say price of car is 0, then you know there might be some issue that data. So, you want to kind of remove those kinds of data points. Similarly, power of the car; if it is a ridiculously small number then you know cards do not come at those powers and so on. So, you can remove those. So, that is another important thing that you will see in this case study and how do you think about out layers the very formal mathematical ways of thinking about out layers and removing them and there are also more common sense notions of what out layers are and then you can remove these out layers.

And, also if you have certain categories with very little frequency of occurrence then you can think about whether you can combine two categories into a same category and so on, so that you have a much more compact data set that you can work with.

(Refer Slide Time: 13:43)



Framework

- Solution conceptualization (contd.)
  - Filter data based on logical checks
    - Price, year of registration, power
  - Reduced number of data
- Method identification
  - Linear regression
  - Random forest

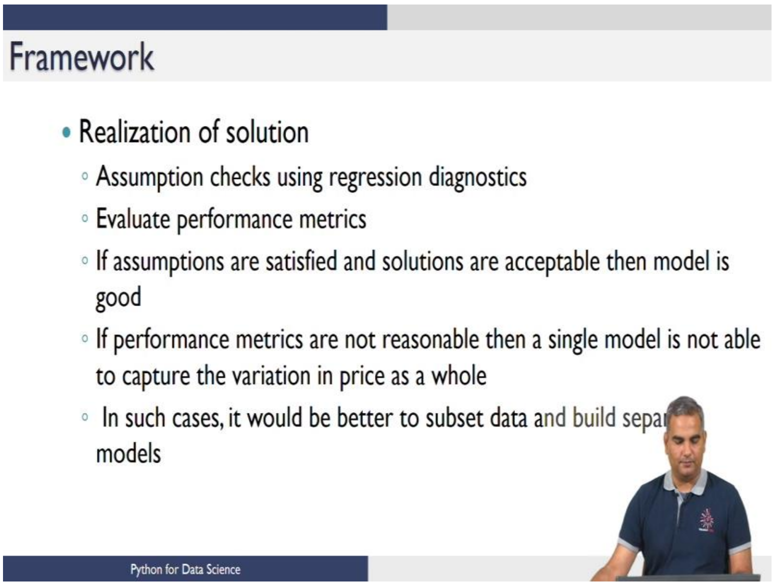
Python for Data Science

So, as I mentioned before, we can filter data based on logical checks, price, year of registration power. So, in some cases you will notice in this data set the year of registration does not make sense at all. So, there are years which are past 2019 and there are years which are very far back from 2019. So, you can remove data like that and then get a reduced number of data.

And, as I mentioned before we are going to look at two different techniques one is linear regression which sees if there is one linear model which will fit the output variable as a function of the input variables. If that is the case then the model is simple, analysis is simple, variable importance everything is simple. So, we will first try that and also we will see whether a non-sample method such as a random forest approach will work in this case better than a simple linear regression approach and basically it is a trade-off in terms of understandability of the model, how much better one model does over the other and so on.

And, what you are going to use this model for; are you going to embed this model in some other optimization and so on, in which case you might want to worry about complexity explainability and all that. Otherwise if you are just going to use this in this example if a random forest model that is much better you might simply go ahead and use it for this example.

(Refer Slide Time: 15:11)



Framework

- Realization of solution
  - Assumption checks using regression diagnostics
  - Evaluate performance metrics
  - If assumptions are satisfied and solutions are acceptable then model is good
  - If performance metrics are not reasonable then a single model is not able to capture the variation in price as a whole
  - In such cases, it would be better to subset data and build sepa models

Python for Data Science

So, you will notice that you will do assumption checks using regression diagnostics. Linear regression models come up with come with certain assumptions which once you build a model you can verify whether the model allows us to kind of understand whether this data follows the assumptions that we are laid out at the beginning. You can evaluate performance metrics.

In this case the performance metric is very straight forward which is basically how much error is there in your prediction and if you are happy with the result you simply live with it; if you are not happy with the result then you might say well there might be some other processing of the data I can do and then maybe get better models. Maybe one idea would be to better subset data and say I will build separate models for this cars of.

One very simple idea is let me group all the cars with very low price and then maybe build a model which I might call as a model for low priced cars and then I could subset all the cars with a larger price and then build a model for higher priced cars and so on. So, these are things that come out of your analysis after you build the first level of data science solution and then understand whether that is good enough or not or you have to do more to make your results acceptable for use.

So, in the lecture that follows this, you will see how you solve this whole problem in Python and look at these various choices and then see how you give a holistic solution to this problem. Hope these case studies are useful in your journey towards understanding data science and in particular using Python to understand data science and I just want to say that this is just a beginning. These are simple case studies, nonetheless they are themselves useful case studies for you to understand a process by which you solve these problems and as you go forward hopefully you will encounter a more complex and rich problems that that you get to solve in this area of data science.

Thank you.