1.5 Transcript

So the rest of this session, we're going to look at ChatGPT plus strategy, plus is the paid version of ChatGPT.

And there's a lot more you can actually do with it than you can do with the free version.

And the main thing you can do with it that's different is that you can actually load images and PDF files and audio files and ask it to work with that and interpret it, and do you know things with that.

The example that I want to show you is to deal with analyzing data and to analyze data, you need to give it the data.

And the data is typically going to be sitting in a file of some sort. So you want to be able to read the file and then do the analysis with that.

And also it gives us a little bit more flexibility and tokens so we can actually end up with a few more, tokens or get larger inputs and outputs into our system.

We can also use a device camera to examine and interpret your surroundings.

So you can actually think of this as something like, you put it on the camera and then it will tell you to go forward, go backward, or maybe, I don't know, like something, a sci fi kind of thing, but pretty soon it will be there.

And you can generate images with OpenAI's Dall-E two, which is kind of helpful too.

So let's take a look at some of the things we can do with it. Let's start by looking at the simplest, which is to load and interpret images.

So here's an image that of a bird sitting on a tree.

And we want to figure out what bird it is and maybe even ask where it is, which part of the world it is in, and ask you to interpret that.

But we can say, hey, can you do that? So here, if I load the image in and then, ChatGPT.

Plus you can load a file in, so I load, I loaded this file in and said what bird is in this particular image?

And it tells me that the bird is a Hoazin and gives me the biological name for it.

Hoazins are distinctive birds found in the swamps, riverine forests and mangrove of the Amazon and the Orinoco delta in South America, including Colombia.

They are actually very pretty birds. So it's very interesting to see them because they have a spiky crest and blue facial skin.

So it tells me all about the bird. So it's done a good job of figuring out what bird is in this particular image.

Right. And so done pretty straightforwardly.

Then I can say, take this image, the bird in that image and contain and give me and create a new image for me that contains this Hoatzin in Times Square.

So now what I want to do is I want to take that Watson and insert it into an existing image or something, or ask it to generate like a fake image for me. Right. So I do that and I say, hey, do that.

And I get the Hoatzin sitting in tight time square.

This is what it can do for us. And I can ask it if I want to make the bird smaller, bigger, make it sit on a man's hand, I can, you know, do all kinds of stuff with it that would make it, you know, more realistic or, if you wanted to fool somebody, you add yourself to the image, you know, all kinds of stuff you can do with it.

So this is a very powerful feature image generation, which can be, which you can use in ChatGPT plus using Dall-E, which is Open AI's software that other software as well that can do this for you, other generative AI tools that can do do this for you.

And, of course there are dangers. The dangers are that people can use this to make fake images and, you know, do all kinds of stuff.

And in fact, you can make fake videos. So there are downsides to this, but it also it's kind of fun.

And, you know, if you want to create like a marketing campaign and you don't have the money to go and invest and, you know, go to a particular place to take a picture and hire someone to do it for you, you can actually just create an image by describing it, maybe a few iterations.

You can get exactly what you want.

So it's a very inexpensive way of, of generating images that you can use in a good way, a useful way, if you want to.

The doesn't have to be all bad. Moving on.

We're going to look at some and spend some time now looking at how we can analyze data using ChatGPT plus.

The main reason for using Chat GPT plus here is because we want to load the data file into ChatGPT plus and, you know, tagged and you need alicense for that or plus account for that.

And you can do actually fairly advanced data analysis. So we'll see. We can do quite a lot of stuff with this.

So this file is contains housing data and it contains information on homes in California in 40 years ago actually.

But whatever. And it has a whole bunch of attributes.

There's longitude, latitude, the median age, these the the data contains homes inside blocks, right.

So it groups of homes and you have the median age of the home, that of all the homes in that block, the total rooms, total bedrooms, the population, the households, the median income, the median home value. So these are all raw data items that we wanted to work with.

We could tell it, hey, can you read this data and run a regression that uses median home value of the dependent variable and median income, population and households? Just these three as independent variables report the R-squared, the intercept, and the beta coefficient.

So we want to know what the r squared is. And you know a regression to return an equation.

So we want to know we want to construct that equation. So we tell it to do this.

And it can do that for us. It runs this stuff here and tells us that the r squared is 0.5125.

The intercept is 41 to 26. The beta coefficients are median income 41,000, population 45.84 and -45.84, and households 140.36.

So looking at this we can see that median income is, you know, a pretty important factor.

Population is a negative factor. So the more populated blocks tend to have a lower median home value.

And households, the more households they are, tends to have a positive value.

So oddly enough. Okay.

And it tells us that the results indicate that for 51.25% of the variance can be explained by the median income, population and households.

So it tells us how much of that. So but what does this mean? I mean do we really know what all this stuff means?

And we can, you know, we have looked at it.

We want to find out, is this a good result or a bad result. So we can ask the chat.

We can ask ChatGPT to actually do some of, analysis of our analysis really, so to speak.

We can ask it to help us interpret the results. So 50% is not bad for r squared but not great.

We can ask it to help us to interpret the results.

We can look at visualizations with plots to see, you know, how good or bad it is doing to sort of get a sense for whether our results are good or not.

Right. And we don't have to write code for this is going to just do that for us.

So we're going to generate three plots observed versus predicted median home values,

residuals versus predicted median home values, and distribution of residuals, and Chat GPT will explain to us how to interpret these.

So it'll also tell us you know whether these plots are, how useful they are, how good or bad they are etc.

And we have to ask it of course, then it'll tell us all that.

So let's say visualize the regression results with plots. That's all I have to ask it.

And what it does is it produces actually a lot of information.

But I'm going to just, extract the three plots from that information and we'll see the plot.

So it produces one plot which is observed versus predicted median home values.

And it tells us that. Why this plot is useful?

Because it shows the relationship between the observed and predicted values.

Now, ideally, what you would like to see is you would like to see this yellow, orangish yellowish thing here, like this one to be more, like more like, you know, well swung upward so that it looks a little bit like this right around this line.

That would be perfect. So this would be perfect. And this is what we're actually seeing.

Right. And, but we can see that there is some kind of relationship, especially in the lower areas.

And, you know, it'd be nice to be a little bit more. And it, it tells us something about our plot here.

The second plot, it shows us the we can ask that you will explain it further.

But, you know, for now, let's just look at that. It tells us the residuals versus predicted median home values.

And it just, displays the residuals that the difference between the observed and predicted values, the difference between the actual and the predicted against the predicted values.

So this will tell us that these are the, the y axis is the differences and x axis is the predicted median home values.

Right. So the y axis is the differences y and x is predicted home values.

These are predicted home values. And that's the difference from the actual.

Ideally one would like to see the residuals to be zero.

Right. Because we don't want to have any error. We want the actual the predicted to be exactly the same.

We want this whole thing to be as close to the x axis as possible.

That would be really nice if we could get that, and the third graph that it shows us is the, distribution of the residuals and the it ideally we want the distribution to be normal.

And, this looks reasonably normal. But you know what?

Let's see what ChatGPT tells us. We can see the model fits the data, but not perfectly.

If you look at the angle of the orange dots, the actual versus the angle of the red line, then we saw here in our original thing here that, you know, it wasn't perfect for the second one, the residuals should be clustered around zero, which they're not.

Right. They're pretty much up here. Right. So we what we would like to see is clustered around here.

And then for the third one we want to see the a normal distribution.

It is normal ish but a little bit skewed to the right here.

So if we look at this here. We see that there's a skew here on the right hand side.

Right. So if it was purely really normal it would be somewhat like this. And it's not like that. So we have a bit of a skew there. So that's not necessarily bad because you're not going to get perfect results.

But we want to know like is it good or bad.

So what we can do is we could for each of these plots we could tell ChatGPT, can you help us figure out what it's doing?

Is it good is it bad? I don't know, you know, what's going on. So I could say, hey, could you use the first plot observed versus predicted median home values to give me a sense for how good my results are in about 100 words,

I'm thinking about 100 words because when I write it the first time, I got, you know, huge amount of results.

I just want a quick summary because I'm lazy. So and we always want something really quick.

So what it does tells us that the observers, the predicted median home values shows a moderate fit.

And that makes sense because it's telling us that the R-squared is 51.25.

Well, many points align around the ideal y equals x, not y equals x nine.

There's noticeable scatter suggesting variability of the residuals, to improve the model it tells us, consider adding more relevant features, exploring non-linear relationships or using advanced regression techniques.

Overall, the model provides a reasonable but not perfect prediction of median home values, which is pretty much what we can see over here.

Right? So it it helps us. Not only does it give us a bunch of plots to help us understand our, our, regression results, but it can also help us interpret those plots.

Right? And it can do all that kind of stuff there for us. So now I tell it,

Hey, could you rerun the regression after standardizing the variables?

So notice that when we looked at our original data over here in this here we are running a regression.

But in our for our data is very, the means are very different.

Right. So if we look at the population, for example, these are all in the hundreds and the thousands.

If we look at median age, that's in the tens. We look at median, households also in 100,000, 100,000 and the total rooms is in the thousands.

Right. So these are very, very different in magnitude, which is not a good idea.

You know, you really want the mean and the variance of each of the distributions to be pretty close to each other when you, each of the variables to be close to each other when you're running a regression.

So what we can do then is we could say, can you rerun the regression after standardizing the variables?

Would we get better results then? So I go back to ChatGPT and it turns out to be too complicated.

You know, it's, ChatGPT unable to do that. In practice,

if you went variable by variable, it could probably standardize it, but it will be very tedious.

You know, you're gonna have to ask, take this variable, standardize it, take this variable, standardize it, etc., etc. would be way tedious, but it could probably do that.

But running all of them together, it just turns out to be too complicated.

So what ChatGPTdid instead is it gave me, it gave me a list of methods for standardizing the data and said, and also gave me the code and, you know, and said you could actually do it with this code over here. So ChatGPT could not actually standardize it for us very easily.

We need to write some code to do the standardization. So what I did is I took the data and I put it into a new file called Standardized data dot csv.

And well, I've done the standardization and I give it back to ChatGPT.

And then we'll see what chat we can do with this particular data here. So now I have this analyzed data there.

And I read it into ChatGPT. And I tell it that the first row of this file contains column names.

Could you split the data into 70% training and 30% testing?

Notice that when we ran our first regression, we didn't have a training and testing.

We ran the regression on the entire data and looked at the results on the entire data.

But typically when you're doing data analytics, you want to be able to, have a separate training and a separate testing data set so that you train the model on the training data set and then test it or report the results like the R squared and the mean squared error, etc. on the testing data set and run a regression of the training and report the R-squared rmse on the testing data.

The dependent variable is median home value. So I'm giving it the full context that I wanted to run.

So ChatGPT runs this stuff and it can tell me now like just like before it does me now that the R-squared is 0.536 or 53.6% and the mean error is 78,000, which is, , it's actually the, the root of the mean of the squares of the errors, but you can think over to the actual error.

So it's wrong by about 78,000.

So he's saying if you get, housing prediction of 400,000, you can expect the range to be about, 322,000 to 478,000.

Right. So that's the range of the error there. It got the average error of the predictions.

So now I have the R square and the average error.

And I want to know, is this any good? So I say, is the r-square value considered good?

I'm asking Chatgpt to believe that. ChatGPT comes back with a long explanation.

It says depends upon heavily on the context and complexity of the data and then gives me, you know, the context in which I'm doing it tells me that social sciences.

Is it something else or economics or whatever? What's the purpose of the model?

Is it for explanatory purposes? Then it might. If it is only for explaining, it might be very useful.

If it's a prediction, maybe not. Because, you know, if you give a prediction, then the error, the error might be, the r squared is not enough to explain the model, and then you might not get good results.

So it's telling us if you just gonna explain something then yes, you know, you want to say median income is, determines housing prices, then yes, you can do that.

But if you want to predict based upon this, then maybe not so great. Then you have to take a simpler model and compare that.

We don't have a simpler model, but I could we could do that. And it tells us, look at the, the variable of the dependent variable variability.

And it might be, if it's very variable it may be hard to do that.

So I could actually ask ChatGPT to take this and analyze it further.

Right. And, and take the dependent variable and do some analysis on that.

But let's skip that for now and we can move on to doing some more analysis with it.

I can say, what do you suggest I do next? And for context, I am trying to predict home values in California.

So I'm asking you for suggestions now.

So it tells me, okay, if you want to do something then these are some of the steps that you should probably use, or I should probably use to improve my model results. And the very first one it says expand your feature set.

It's saying that you don't have enough, you may not have enough features.

So since location is a critical determinant of home prices. And notice that now because I have said I'm doing home prices in California,

ChatGPT is using its knowledge of homes essentially to offer home pricing to add some flavor to its recommendation that it says location is important.

So it says you should do that. It will include more granular location data such as zip codes, neighborhood indicators, schools, parks, commercial centers, these kinds of things in my analysis.

Right. So if I could get that data, then it's saying that your results would be better because ChatGPT knows that location is an important thing.

Economic indicators, local employment rates, median household income, the economic growth indicators.

These are all important things in figuring out house prices.

If the neighborhood is, like has a low employment rate, then presumably housing prices would be lower.

So you want to actually include various economic factors for each locality inside the in the database in your data as well.

Then you also want to look at the physical characteristics of the home, its age, condition, style, whether there's a pool, renovations, all these kinds of things.

We can't, since we are dealing with blocks of homes. Maybe we can't really do this kind of data, or we could try to find some kind of average renovated expenditure per home on renovations or something like that.

But, we already have age, so here, ChatGPT is kind of messed up a little bit, but, it's essentially telling us how we can improve our data so that we get better results.

Right. And then you can use advanced modeling techniques. They've given a whole list of them for us.

And then GPT tells us that we can improve the data quality, do some more advanced feature engineering.

We, like, create polynomial features or aggregate features to see if the predictive power gets better.

And this is sort of tied up a little bit with the advanced modeling techniques, because we can use things like Ridge or Lasso to figure out which features are more important, which are less important, and then sort of do more, engineering on them.

So there's a lot more you can do, and it will keep telling us, you know, giving us ideas on what to do, scaling and normalization, ensure they're properly scaled normalize.

Again, we have to look at graphs, look at averages, look at variances to see if that's the case.

Look for market trends. Maybe go out and do some surveys and figure out what's going on, or talk to experts in the domain so that they know we can get better ideas for, doing our thing there.

Now, since we can't do all that in the next few minutes, we will just go ahead and see if we can do other kinds of analysis with this.

So I'm going to say use the same data split, the 70% and 30% and then do a random forest,

do a random forest model on this and then draw a graph that shows feature importance, because we want to see, you know, which features are important.

And we can ask you to do a random forest, the random forest algorithm on it to figure out what's more important and what's less important.

So I run that, and it tells me that the most important feature is median income, which is here followed by people per home followed by density, which density is a number that tells me how many, homes are

There in a particular block. Really.

How many, what is the number of homes and households inside of a block?

Number of homes, number of households inside a block.

So what this does is it illustrates the importance of each feature and the random forest model and feature with higher bars.

It tells us exactly how to interpret this, and you can focus on the more relevant features when further refining your model.

In our case, we don't really have very many features, but sometimes you have hundreds of features and you don't really want to use all, you know, like say 150 features.

So you can use this kind of thing to figure out which ones are important, and then rerun the model with only the important features, because partly you want to ensure that your model will work out-of-sample and, you want to drop unimportant features because then you have less likelihood of fitting, overfitting your data.

Then I can ask it to actually do some kind of, feature engineering for me so I can say, here, take create a new data column and we can call it value category.

And the idea here is to compute the median of the median home value column and then replace every value with a one if it's greater than the median and zero otherwise.

So what we're doing is we're creating a categorical variable which can be either one or a zero.

And, we're going to use, going to predict whether a home is a high value home or a low value home.

So if it's a one, then it's high value.

And if it's zero, it's low value.

That's the idea here.

So we're going to use a different kind of prediction mechanism using categorical prediction rather than using continuous variable prediction.

So I do this and it does that. It says hey I've created this value called value category column.

And it explains exactly what it's done using the median home value, etc.

Right. And very helpfully, ChatGPT has also highlighted bolded the two things that are important here, the value category and the median home value.

Right. Which is kind of nice. Then you can say, can you run a random forest classifier with value category?

That is the, our new new created column as a dependent variable and use the same 70% training, 30% testing split.

And then we want to report various, metrics on classification models over there.

And of course, you want to make sure it doesn't include this because this would be very highly correlated with the value category column.

So we we want to drop that with the median home value because that's that was our original regression prediction.

So we tell it to report the accuracy the precision the recall and the area under the curve for this.

And it does that for us right. It tells us it's a 85.89% accuracy, 87.12% precision, 84.5% recall, and then 93.66% area under the curve.

if you're familiar with these terms, you'll understand what they are. But the basic idea is very simple.

Accuracy tells us how accurate our model is in predicting whether something is high or low value.

What percent of the times does it get the right answer, basically.

Precision tells us that if it's predicting something that's a high value, then how often is it correct?

So if you take all the high value predictions and then look at the ones that are actually high value, then 87% of the time they are actually high value.

So when it's saying something is high value 13% of the time it gets that wrong.

Most of the time they are actually low value. But our model is predicting high value recall tells us if from all the high value homes that we have in our data, how many of them have we predicted as high value?

So we found 84.5% of all the high value homes.

So, 16% or 15.5% of the high value homes were incorrectly predicted as being low value.

That's what this is telling us.

And the area under the curve is just a metric that tells us how good our model is in differentiating between high and low values.

And this tells us it's 93.66, which is actually quite high.

So it's doing a really good job of this prediction And that's what ChatGPT tells us.

The high AUC score suggests that the model does a good job at ranking predictions rather than merely classifying.

It's effective at distinguishing between the two categories. Right.

So it tells us our model is pretty good. So then I wanted to draw the ROC curve.

The ROC curve is just another way of looking at the AUC. And it can draw the curve for us.

And you already saw that ChatGPT is good at drawing curves and you know, all those kind of things.

So it just draws a nice little curve for us. And this tells us that essentially, you know, this space over here tells us that we have done a good job of prediction because there's very little, that's not inside this.

The higher this curve is, the better your model. That's basically 94% of this.

This is 94%. The space under the curve is 94% of everything in this box here.

Right. So that's the box that we're looking at. So in short, we can do a lot of analysis with ChatGPT, or ChatGPT plus anyway.

And that is, typically you would write a lot of code to just do exactly what we asked ChatGPT to do for us.

But you don't have to do that.

You can ask ChatGPT to do it, and it'll probably do that as long as you are very specific in your instructions and you tell it exactly what you want.

And you know, again, remember, it's a question of giving it stuff, getting results, asking you to explain them, refining your, refining the, the thing that you want, telling it your context again, maybe giving it some more information and trying to get it to do exactly what you want.

And we can see that it works pretty well here.