ADS-506: Applied Time Series Analysis

Lab 3.1: ARIMA Models and Predictors

Hi, everyone. Welcome back. So in this lab, what we're going to do is we are going to discuss ARIMA models and how to determine your ARIMA order, the AR, the integration, the MA, or the PDQ variables, how to determine those. We're also going to look at integrating regression models, so it's an ARIMA regression. And yeah, and have some fun with that. So, all right. So I'm going to use this data set. This data set, it comes from Kaggle. I've already shortened it up and cleaned it up a bit, but it's bike rentals from a shop in Seoul, Korea. They did it hourly, but like I said, I cleaned it up a bit just for purposes of the video. All right. So let's load in our libraries, and then I'm just going to plot our data set. I'm just going to take the... Actually, let's go from here, before I get ahead of myself. All right. Let's load in our libraries, load in this data set and then take a look at it.

So we can see here, we have dates. We have some continuous variables, temperature, community, wind speed, rainfall, and then the number of bikes rented. Right. So number of bikes rented, that's what we're interested in. That's going to be our outcome variable. So let's just plot that and take a look at that by itself. So let's just look at bikes rented. Now, since we're dealing with calendar days, right, I just assume, and this really is an assumption. There is probably going to be a weekly seasonality, right. So meaning. I think bikes rented on Monday is going to be similar to what bikes rented last Monday were, and the Monday before that, and the Monday before that. Bikes rented out Saturdays would be similar to what last Saturday was, the Saturday before that, and Saturday before that. So I'm saying my frequency is going to be seven. I don't have to say a frequency. I can just turn this into time series and just figure all that out later, but I think it's safe to say.

All right, so I'm going to plot it because that's what you do. The first thing you got to do is you got to plot your data, take a look at it. And right away, I'm noticing some really sharp drops. So one thing in this data set that I don't have here, but if you went to Kassel, you downloaded it, you would have other, you'd have some categorical variables like is it a holiday? Is it a functioning day? Is it winter? Some stuff like that. So that might explain some of these drops, but for this purpose and this lesson, we don't care. So what I might do if this were a real analysis, and these were the only, these elements that I had, I would probably impute some of these values. Now, dealing with time series and you impute some values. It's a little bit tricky, right. Sometimes you can't just take the mean of the value before and after.

It depends on what it is.

In this case, you probably could. You'd probably take the mean of the bikes rented the day before and the day after your missing values, and then put that in there. You should be fine. You just got to be careful with time series, because time is a factor. And so it's not as easy. You definitely don't want to remove them. You don't want to delete this stuff because like I said, time is a factor. And you start deleting it, all of a sudden you're missing some days. And then that throws off your whole analysis. So just be careful if you're going to be imputing that. Okay. All right, so that's what it looks like. Now, we can take a look at this. Let's just look at the ACF plot right away to determine what, if any correlations there are. So let's do bikes, auto correlation. All right. So let's take a look at that, generally determine. Now. What you might notice here is we have .51, 1.52. That's because I set the frequency, so these are actually seasons.

So in the first season, second season, or first week, second week, third week, but let's actually go out a bit more. So we have three and a half weeks. So let's look at maybe seven weeks. All right. So we can see here, there is a slow slope down, so it would suggest trend. But we also see this oscillating pattern, right, and we could see kind of a spike every season, so on the season mark. So this spike would suggest seasonality as well. So now, there's a couple ways to deal with that. Right. So what we could do is we could take our bikes and then we can make a difference, just one difference. And then I'm going to make this... Let's look at the ATF plot for the difference. All right. Oh, too far. Let's look at 21.

All right. So we're looking at three weeks here. So I just took one difference. That's what difference, that I just took one difference. I could take a seasonal difference and that would just be different bikes and then lag seven. That would just take basically the Monday subtract from Monday before, and then that value and so forth. So I just took one. Basically I took the difference, so December 2nd, minus the value at December 1st, and then that gives us the value. So we can see here, the ACF, just taking one difference. We don't have a seasonal pattern anymore. What I'm looking for is I'm looking for spikes sticking above and below the blue dashes, especially on the seasonal part. So there's none on the one, the two, the three, there's none really there. Nonseasonal components, I do see one here and I do see one there. So that's one, two, that's three.

So now, this is the ACF plot. The ACF plot is used to determine the MA part of our ARIMA model. So now if you remember, should remember, should have done your reading already. Your ARIMA model is this, PD and then Q. This will be on the quiz. So this really translates to ARP and then we have an I. Don't

really care about it, doesn't get no love. And then we have our Q, or the MA. Sorry. Q. All right. So the ARs are auto regression. The I is basically the different theme and the MA that is going to be the moving average, which is the correlation, not the correlation. It's going to be the relationship, the error terms have with previous lags. So to determine these, we will look at certain plots, right. So the ACF plot that determines the MA part. Okay. And then the PACF part determines the AR part. So just remember that. It'll be on the quiz. Actually, I don't know. I'm just going to assume it's going to be on the quiz. I would quiz you.

All right. So we already took the first difference. So the I of our model, this is a one. It took one difference. Now, looking at the ACF, this is going to determine our MA, so we don't have any seasonal ones. So I'm looking for non-seasonal and I can see one, maybe two, I mean, sorry, one or three. What I'm looking at is the number of lags. This first one starts at zero, so we're not counting that one. So it's 1, 2, 3. So let's just say MA three, just for giggles. All right, so now we're left with the AR part. So that is PACF, right. So since we're looking at the difference, I have to look at the difference again. Wouldn't make sense if I looked at the normal one. All right, so there's still some differences here. So we can take a second difference and determine that. If I take another difference, then this one becomes a two, just because I did the difference again. Even if I did a seasonal difference, right. If I took a seasonal difference and did them one time, that's only a one.

If I took a seasonal difference and then a one lag difference, that's two. Two differences. If I took a one lag difference and then another lag difference, we're still at two. This number is just the number of times you did it, not what you did. So, all right, so I just took... I'm just going to stick to first order of difference. All right, so we have a 7, so it's 1, 2, 3. This 3 sticks out, so I would probably say this is going to be an AR 3 just because this lag is higher than all the others. Could be an AR 1 or an AR 3, but let's just stick to AR 1. I'm sorry, AR 3.

All right. So now that is pretty much how we are determining our ARIMA models. Now, it's not an exact science, it's still open to interpretation. But we do have some statistical analysis backing us up, or we're not just pulling stuff out of places. We're not just saying, all right, this is an ARIMA of 7, 1, 8 or something that's... That wouldn't make sense. So we're looking at the PACF and we're like, all right, so that's 1 definitely sticks out. And then just and this one sticks out, so that's why I chose a 3. So because that's one, two and three. So you could say four, but just looking at it, I would still say 3. Just a hunch, I guess.

All right. So let's make our model. All right. So my ARIMA is going to be

ARIMA, and I'm interested in bikes. And I'm going to have to specify the order. We can see from the tool tip. X, that's our object. That's the object of interest. We are interested in bikes, the order, that's what I'm tying it. And I'm going to give it an ARIMA 3, 1, 3. Now earlier, when I was talking about seasonal, right, the seasonal difference or seasonal component, we can put that in here as well. Right. So if I saw a spike at season one, season two, season three, then in my ARIMA model, when I would just put that under seasonal. And it would look just like order. So let me just put order here, first. Order, so I'm saying this is a 3, 1 3. I'm saying this is a 3, 1 3. All right. If we wanted to add the seasonal component, we could just do that, and we can say it's zero. We didn't take a seasonal difference and then a one. And so that's how you would do it.

So, all right, so let's run this. And then what I'm going to do is look at the summary of my model. Let's run that. All right, cool. So what we see here is we see our coefficients, right. So I said, this is an AR, I'm sorry, an ARIMA 3, 1, 3. So we have three AR components and three MA components. The summary gives us our root means squared error, and also gives us the AIC. So both of those are metrics you can use to determine model performance, or which one's better. So if you're curious about what those coefficients are, those are basically your back terms. Right. It's like a regression model that just creates a formula, an algorithm, same thing with ARIMA. So at this time we have our object of interest, our outcome variable, the thing we want to know, which is YP. So that is the value of Y at time T. In this case, would be the number of bikes rented at time T. Right. And so then we have that. And then we have some sort of constant, so that's B not... I just can't type.

And then we are going to have our coefficient, which is actually the first AR term is going to be a negative .7, times the value at YT, minus 1. And then the second one is another negative. So that's be .18, times the value of Y at T minus 2. And that third one is also negative. Nice. It's practically zero, but we'll put it in there anyways. .007, times, Y at times T minus 3. So we are saying that the outcome variable is equal to some constant, plus the coefficient times the value at the previous time, plus some coefficient, times the value at the time before that, plus a coefficient times the value before that one. Right. That's just the AR part. The MA part is basically the same thing. So these are all negatives too. So, all right, let me just... This will be, I'm going to stick with O, because why not? But this time it's going to be an error term at time T, minus one. This is .07. This is going to be, the second one is .3, 4 times the error term at T, minus two. And then this last one is 3.3, 1 times the error term at T, minus three.

So that's what this ARIMA model's saying, actually we're taking a first difference. So it's actually Y prime. So all these are going to be Y primes.

That just means the value of Y, minus the Y before. All right. So that's what this ARIMA model is predicting. So if we want to see what it looks like, then we can just do the forecast, right. Forecast this model, the next 12 terms, or next 30 days, and it would give us a forecast of where we can expect. So that's just a real quick overview on ARIMAs and how to determine, especially your AR component and your MA component. So you're going to look at your PACF plot to determine your AR component and your ACF plot to determine your MA component. Sorry, hold on, confused myself. You're going to look at your ACF plot for your MA and your PACF for your AR. All right. So now, this is cool and it's actually pretty effective for determining your forecasted values. And it's actually pretty good considering that we don't have any outside predictors. Right. I just have bikes here.

But now what if we had outside predictors? What if we had things that we know influenced the number of bikes rented? So what about predictors? All right. I'm glad you asked. All right. So let's take a look at our data again, so had... Right. So we have these. We have temperature, humidity, wind, rain, right. And we're saying that those have an effect on the number of bikes rented. So let's put that to the test. All right. So first thing I'm going to do, right, because I'm going to make a model, regression ARIMA model. I'm going to make a training set and I'm going to make a test set. So training set, this is sole. Now, I have a year's worth, I actually, I have 365 days. So what I'm going to do is I'm going to take the first 335 rows, or take all my columns. And I am going to actually, I'm going to remove my dates. And yeah, I know I could put it in here. That's fine.

All right. So the reason I did the first 335 days is because this is my training set. I'm going to use the first 11 months to train my model. And then I'm going to test it on the final month. In other circumstances, when you make your training set, you would probably take a random sample, a random split, 80/20, 70/30, something like that. I can't really do that with time series, because like I said before, time is a factor. So you have to take that into consideration. So that's why I'm taking it in order, I'm taking the first 335 values. Right. And then my training set, I'm sorry, my test set will be the last, so 336 to 365. And again, I'm just removing the date column. And in case you're wondering, yes, I could have just removed it inside the brackets, but just didn't want to. All right, cool. Now, what are we interested? What do we want to know? We want to know the number of bikes rented. That is our outcome variable. All right. So let's do this. So outcome variable, and that is going to be our training bikes rented.

All right. So now, real quick side note, we are going to use our predictive variables and I'm actually going to call them predictors, right. However, I can't just select them. I need to put them in a matrix. Nothing's really

changing, it's just the way the ARIMA function works. It needs them as a matrix. So training set, and I want the first four columns. Okay. So I'm going to get these, and so now they exist in my environment. All right. So let's look at two functions. Let's look at the auto ARIMA, right. Let's say you need to come up with something quick, feeling kind of lazy, or you don't really know what to do. You can do the auto ARIMA. So this is the auto ARIMA model, and this will try to come up with the best model. It doesn't. It comes really close. But I would say that this is a good way to start. It's a good starting point. You would still want to tweak it, finesse it a bit and I'm sure you come up with a better model, but this is probably a good starting point.

All right. So now, look at your tool tip. So auto ARIMA, Y, that's our object of interest. Our object of interest, the outcome variable. The other thing in here, actually it's not in here, is predictors. We can put predictors in here. We can tell it we have an object of interest, it's bike rented. That is a time series object, but we also have these predictors and these predictors can influence the number of bikes rented. So we will use the xreg. Okay. So the cool thing about RStudio is that it does give us tool tips and hints and help. So it's just telling us the xreg and what it's used for. So I want that, and my xreg is going to be equal to my predictors. Let me just write my summary.

All right. So let's run that. And this comes up with an ARIMA 1,0,4. So we have one AR component, four MA terms, right. And we have our Y intercept and temperature, humidity, wind speed. So what's it saying is we're going to start off with 788 bikes rented. Just right off the bat, whatever given day it is. Now, temperature for every one unit increase in temperature, we can expect an increase of 24 bikes rented. However, for every unit increase in humidity, we will subtract 6 bikes rented. Right. No one wants to ride their bike in the humidity. And if it's windy, apparently if it's windy, we're going to increase the number of bikes rented, helps you paddle faster. Along with that, we have our ARIMA. So again, the function I wrote above, right here on line 44. Right. So we have our constant and then we have auto correlation to the previous value. And then we have a relationship with the error terms, the previous four error terms. And all those get included in our model. And it creates a really long equation.

All right. And the other thing we want to look at is the RMSE. So our route means squared error is 224, the AIC is 4,600. Right now, those numbers don't mean much, because we're not comparing it to anything. But if you wanted to figure out the root means squared error, just keep saying root means squared error out loud, because that actually tells you what to do. So the root, which is square root, right. So root mean, just type in mean, and then root mean squared errors. So what are the squared errors? So we have our auto ARIMA model, dollar sign. And we're going to get a very specific

component out of our model. We are looking for residuals. The residuals, it's just another way of saying errors. The residuals is the difference between the fitted values and the actual values or expected values and actual values expected over observed, or it is the difference. It is the errors. And we got a square, so root mean squared errors.

So we do that and we come up with T 24, 3003, which is what they have. So our math check up. That is just a helpful hint. Okay. All right. So we did the auto ARIMA, right, and it came up with a value on its own. But I think we can do better. Now, the reason I think we can do better is because we already did some of the work. Going back up here, right, we already did this ARIMA model right here. And if we look at that summary, that RMSE is 287. So the auto generator one is 224. So the auto generator one is better, but their model is a 1, 0, 4. Our model is a 3, 1 3, so let's put our order in. So our turn.

All right. Let's do this, so this is my ARIMA model. And so this is going to be capital A, ARIMA. And what is that thing that we're interested in, what is Y? That's going to be our outcome variable, so outcome, right. And we have to specify the order, so the order and that is going to be, we said a 3, 1, 3. 3, 1, 3, and now we're going do the xreg, right. And so these are going to be our predictors. So let's look at this summary, and let's see if we can beat the robot. All right, let's run this. And the RMSE we got close, the RMSE for the auto generated one was 224, 300. Ours is 224, 82, so not better. RMSE is a little bit higher, but not by much. And ours is, so we said it's AR 3, one difference, MA 3, right. And so these are the coefficients for those terms. So our model is not better, but it's a fun exercise. Maybe if we took a seasonal order to it, we could probably improve. So it just depends with the speed of the processors now.

We can run a ton of different models and then settle for the one, what the lowest root mean squared error, or the lowest AIC. I use the lowest AIC and the lowest root mean squared error. Now, when I look at the root mean squared error, that's telling me how far off I can expect to be. Right. So I can expect on average to be off by 200, so 200 bikes rented. Now, there's going to be other factors that are obviously weighing in on the number of bikes rented. I'm only looking at four, but the actual data set had, I think 10 factors. Some of them were categorical factors, categorical variables. So I'm just looking at a little snippet, but for forecasting purposes, we're pretty good. All right. So we have models, so let's see how good they are.

So forecasting. All right. So now, remember when we had our predictors, we termed them as matrix. So we are going to take our test set, and we are going to use our test set because this is data that the computer has not seen.

And we're going to test how well, or how accurate our forecasts are. So let's get our predictors and then set it as a matrix. And this will be test set. And again, just the first four columns. All right, so then here's my forecast. It will be forecast. All right. So forecast, the first thing we got to put in there is the model. So I'm going to go with the better model. Saddens me, it's not mine, but... All right. So the auto ARIMA model, and then we have to tell it what our predictors are. And so this can be the my predictors. Let me go click this. All right. So here we go. And we're done. I'm just kidding.

Okay. So these are our forecasted values. So using the predictors that we have in our training set. Right. So the very next day after our training set, temperature was 8.6 and then it rose to 10.7 and covered around 11 and 12. Humidity was a certain amount, pretty windy, no rain, except for maybe that day, and the actual bikes rented. So we're actually going to use that to see how we performed. So it's a auto plot. And I'm just going to take the original data set. And I'm just going to color this red, because red stands out. All right. So then auto layer. We're going to do it with my forecast. Right. Huh. That red X through me off. Sorry, [inaudible 00:40:48]. Okay. So okay, let me zoom in a bit. Let's see here. Then we can talk about it.

Let's go from X limit, 300 to 370. And let's make this one a little bit see through. All right. So what we have here now, the red is our actual, these are the actual bikes rented. And this blue line, that is what the auto ARIMA function said was the better model. And we can see it definitely follows the same trajectory, okay. So we even have some of the same dips as the actual bikes that... Now, there's some extremes, right. We all know what these extremes are. They might be one of the other variables I didn't include in the data set, but we do have that. But overall, it definitely follows the same path. And so I would say it's a good model. And these colored ribbons that you see, these are the 80 and 95% confidence interval. And as we would expect, most of the time, it's within the 95% confidence intervals. There's just two, maybe three periods where it's not. So, but most of the time all the points are within. And so we're good. So this is a good model.

We could probably improve on it. How much more? I don't know if it's even worth it. Right. So if I... Let me just look at the summary. All right. So let's say I put in other variables, and let's say we reduce the RMSE to 220. So we can see it. It's a noticeable difference at numerically, but on the graph, it might not be a noticeable difference. So then you ask yourself, is it worth it? Is it worth the extra effort? Do you got to change some things? Is this an automated algorithm? What work goes in there to... What do you have to do after you change it? So if the RMSC dropped to a 100, actually that's way too crazy. Let's say the RMSC dropped to 200, right. Okay. That's a noticeable difference. That's at least a 10%, 12% different. So then it would probably be

worth changing your algorithm to the new model. So, cool. So, thank you. That is our lesson. And I will see you next time.

(outro music)