Abstract

The object of this project is to predict humidity by time and other weather information to avoid economic lost due to series weather in real life. In this paper, we will build the model with the data from Kaggle to figure this out. We will use Humidity as response variable; use Date, Temperature and Wind Speed as explanatory variables. ACF and Decomposition are used to evaluate the data. Also, we use correlation matrix and linear regression model to find the relationships between response variable and explanatory variables. Finally, we use the regression with autoregressive, Holt-Winters method and ARIMA model to predict the near future humidity. We obtain some well-performed models in this project, and we find that Temperature is the feature mostly affect Humidity. Wind Speed does not have a strong relationship with Humidity, but Wind Speed will improve models since it carries some useful information. The multivariate polynomial regression model is the best model we get if we have additional weather information other than Humidity itself. The ARIMA is the best model we get if we only have Humidity and time information.

Introduction

This project emphasizes how time and other weather features influence Humidity. It can be split into several sub questions:

a) How Temperature and Wind Speed affect Humidity?

b) When I have already detected the features of the data, is it possible to do the forecast?

c) How can we determine the parameters p, d and q to best fit the model? Is the forecast what we really expect?

Method and Materials

1. Data

The dataset is from Kaggle (https://www.kaggle.com/budincsevity/szeged-weather). There are 96453 observations with 12 variables; 3 of the variables are categorical variables and the rest are quantitative variables. Our goal of this project is using selected explanatory variables to predict response variables, Humidity.

By checking the correlation matrix (plots 1.1 and 1.2) of the dataset, we find that Temperature and Wind Speed are highly correlated with Humidity, while Temperature and Wind Speed are not highly correlated. Therefore, we select Temperature and Wind Speed to be our explanatory variables. What's more, since if we use all the 96453 observations, which will be computationally expensive in following steps, we decide to only use data of August 2016, which has 719 observations.

2. Individual Variables

To discuss individual variables, we will use ACF, Linear Regression, Residual Analysis, Detrending and Differencing, Smoothing and Decomposition to check if there is a seasonal pattern or a trend.

Firstly, for Humidity, we can assume that there's a seasonal pattern by observing the ACF plot (1.3), and almost all lags are significant. It also contains some up-and-down trends. Then, we use Decomposition to check if humidity has a seasonal pattern or a trend. From Decomposition plot (1.4), we find that there is a strong seasonal pattern and a up-and-down trend.

ACF plot (1.5) and Decomposition plot (1.6) of Temperature are similar with that of Humidity, and the large difference is that Temperature has a up-and-down trend, which is the opposite of Humidity.

ACF plot (1.7) of Wind Speed shows that lags are generally decreasing and almost all of them are significant. From the plot we can find that the Visibility is usually around 0 km/h to 20 km/h, and there is some trend. The Decomposition plot (1.8) shows that Wind Speed has some seasonal pattern, and an up-and-down trend. The trend of Wind Speed has a slightly similar tendency with that of Temperature.

3. Relationships between variables

We will use correlation matrix and simple linear regression to find the relationships between variables.

The correlation matrix (1.1) shows that there is a strong negative relationship between Temperature and Humidity. By checking the scatterplot (1.9), it is obviously that there's some negative linear relationship between Temperature and Humidity.

The correlation matrix (1.1) also shows that there is a negative relationship between Wind Speed and Humidity. However, the scatterplot (2.1) shows that there might not be a linear relationship. It is obviously that almost all points are at the left side of the plot, and that denotes that we probably cannot find a clear linear relationship between the two variables.

4. Models

We will use linear regression, autoregressive regression and ARIMA model to predict Humidity.

a. Regression

Firstly, we use Temperature as response variable to fit a linear regression model. The linear regression model performs well and gives an adjusted R-squared value 0.8129 (plot 3.1). The coefficient is significant. The studentized residual plot (2.2) shows a curvature structure, which indicates the model should be more complex.

To improve the model, we add Wind Speed as explanatory variable. The multiple linear regression model performs better than the previous model, and it gives an adjusted R-squared value 0.8132 (plot 3.2). Both the coefficients are significant. However, even though the adjusted R-squared value increases, the studentized residual plot (2.3) still shows a curvature structure, which indicates that the model is still not complex enough.

Then we try to fit the explanatory variables to a polynomial model. This model works better than the previous multiple linear regression. It gives an adjusted R-squared value 0.8268 (plot 3.3). The two coefficients of Wind Speed are not significant. However, if we remove Wind Speed in this model, the adjusted R-squared value drops to 0.8264 (plot 3.4). Therefore, we keep Wind Speed in our model. The reason why this happens may be that even though Wind Speed is not significant, it still contains some useful information.

From the previous part, we find that there's no evidence showing that there's a linear relationship between Humidity and Wind Speed. Therefore, we do not fit a linear regression model for Humidity and Wind Speed.

b. Regression with Autoregressive Errors

We will conduct a DW test with $H_0: \rho = 0$ v. s. $H_\alpha: \rho > 0, \alpha = .05$ for every model to check if autocorrelation is 0 or not.

Firstly, we apply DW test to our linear regression model, which takes Temperature as explanatory variable. DW statistic of this model is 0.13645 < d_L (plot 4.1), which denotes that we reject the null hypothesis. Therefore, we need to transform data and fit an autoregressive regression model. The new model performs well: it has an adjusted R-squared value 0.7519 (plot 4.2). The studentized residual plot (4.3) shows no pattern and is detrended. This indicates that linear model is adequate. The new model has DW statistic 2.119 > d_U (plot 4.4), which denotes that we fail to reject the null hypothesis, and we can ignore autocorrelation now.

Since the studentized residual plot shows that linear model is adequate, we add Wind Speed as explanatory variable to improve model. The DW statistic of the multiple linear regression model with Temperature and Wind Speed as explanatory variables is 0.13998 (plot 4.5). The new model performs better, and it has an adjusted R-squared value 0.7534 (plot 4.6). DW statistic 2.1182 > d_U (plot 4.7), which denotes that we fail to reject the null hypothesis and autocorrelation is 0. The studentized residual plot (4.8) also shows no pattern and is detrended. We may need more variables that carry more information to create a better model.

c. ARIMA

ARIMA stands for auto-regressive integrated moving average. It's a way of modelling time series data for forecasting. After smoothing the data, we could a draw a line through the series tracing its bigger troughs and peaks while smoothing out noisy fluctuations visually shown in plot 5.2.

The next step is to decompose the data (plot 5.3) so that we can detect the seasonal and linear trend of humidity. From the plot, we find that it has a clear seasonal trend and relatively clear linear trend which tells us that the humidity reaches the top at the end of 2018.

Then, we need to test if the data is stationary because we could only use ARIMA model when it is stationary. From the result of ADF test (plot 5.4), we find that Dickey-Fuller = -3.658 and P-value = 0.02711. P-value is small but not that small so that we may not reject the null and accept the alternative hypothesis that the humidity data is stationary.

Next, we should determine the parameters of ARIMA model. At first, we start with the order of d = 1 and re-evaluate whether further differencing is needed. Then, we use ADF test (plot 5.5) to test the stationary again and we find that Dickey-Fuller = -5.7047, P-value = 0.01. Therefore, we can accept the alternative hypothesis that it is stationary.

What's more, we try to use ACF and PACF (plot 5.6) to determine the q. From the plot, we detect that there are significant and constantly decreasing auto correlations within one day from our hourly data and beyond. And from PACF plot, we find that partial correlation shows a significant spike at lag 1 to lag 4.

Finally, we use the auto.arima to automatically fit the model with the ideal parameters with the seasonality to be FALSE. In this way, we can specify non-seasonal ARIMA structure and fit the model to deseasonalize data. Parameters (1,1,2) suggested by the automated procedure are in line with our expectations based on the steps above. Using the ARIMA notation introduced above, the fitted model can be written as

$$Y_{d_t} = 0.9102Y_{t-1} + 0.8232e_{t-1} + 0.0940e_{t-2} + \epsilon$$

AR(1) coefficient p = 0.9102 tells us that the next value in the series is taken as a dampened previous value by a factor of 0.91 and depends on previous error lag.

Furthermore, we apply more complex model in the following step, compared to the previous analyses. We fit an ARIMA model with temperature and wind speed. According to the

AIC values, we choose the ARIMA (1,1,2) model with two predictors, which are temperature and wind speed. Then we use this model to forecast the humidity of August 2016. There is a clear pattern present in ACF/PACF (plot 5.8) and model residuals plots repeating at lag 24. This suggests that our model may be better off with a different specification, such as p = 24 or q = 24. The light blue above shows the fit provided by the model with 95% confidence interval. We then improve our model by changing our p value to 24. the coefficients of ACF and PACF are in the range of 2 standard deviation. The plot shows there is no strong autocorrelation exist.

Then, we display the pattern of the residuals from the fit by using auto.arima. We find that both ACF and PACF show that the autocorrelation of lag 24 is significant. Since we are using the hourly data, 24 hours a cycle really makes sense. Therefore, we change the q to be 24 in order to comply with the result that we find. Then we display the pattern of the residuals again (plot 5.9) and we find that almost all the autocorrelations of each lag lie within the 95% confidence interval which means we almost have a good fit with the order of (1,1,24).

Next, we try to reserve a portion of our data as a "hold-out" set, fit the model and then compare the forecast to the actual observed values. From the plot 6.0 of the forecast of no holdout, we can see that the forecast values are close to the actual values. But more are the values we hold out, less degree of accuracy we get from the model.

Finally, we pull up an autocorrelation plot for the residuals of the forecast values and we find that they are all within the confidence interval meaning that any of the lagged autocorrelations are not significant. Also, we test it using Ljung-Box test (plot 6.1) and P-value is 0.7218 greater than 0.05, suggesting that there are no significant autocorrelations between successive forecasting errors. To determine whether the errors in our ARIMA forecast are normally distributed with a

mean of 0 and constant variance, we compute the mean and plot the histogram of the residuals, suggesting that our errors follow a normal distribution.

d. Holt-Winter Method

We also use Holt-Winters method to smooth the data and forecast. Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point. From the alpha = 0.93 and beta = 0.65 (plot 6.2), these are both high which telling us that both the estimate of the current value of the level, and of the slope b of the trend component, are based mostly upon very recent observations in the time series. This makes good intuitive sense, since the level and the slope of the time series both change quite a lot over time. Then, we use ACF to check the autocorrelation of the residuals. However, not all the lagged autocorrelations are within the 95% confidence interval, and P-value is pretty small compared to 0.05 which means that the errors are significant. As a result, we think ARIMA is a better model for our data.

Result

ACF, Decomposition and Smoothing techniques are used to check if there is any trend and seasonal patterns. Linear Regression models, Multivariate Polynomial Regression model, Autoregressive Regression model, Holt-Winters method and ARIMA model are used to predict Humidity.

After applying the linear regression, the studentized residual plot (2.3) tells us that we need a more complex model. Therefore, we fit a multivariate polynomial regression model. This model offers the best result in predicting humidity when other weather features exist.

ARIMA model also performs well. After examining, decomposing and smoothing the data, we finally determine the order of (1,1,2) from auto.arima (plot 5.9) and change the q to be 24 so that any lagged autocorrelation is not significant. Then, we forecast the humidity by hold-out testing the data of the last 15 hours (plot 6.0) and we also use the data to forecast the humidity in the next 24 hours in plot 6.4. Also, we check the residuals of the forecast values in order to test if it is significant and normally distributed. In comparison with the Holt-Winters method, we find that ARIMA is a better model after using Ljung-Box test and ACF to check the residuals.

Discussion and Conclusion

For the ARIMA model, we find that after we change the order to be (1,1,24), the hold-out forecast is less accurate when we hold out more actual values. Maybe we should try another prediction model, such as SARIMA.

Back to the beginning, we can answer those questions now. Humidity and Temperature has a strong negative relationship, while Humidity and Wind Speed are not highly correlated. We can do the forecast after fully detection to the data. The result is good and absolutely what we want. However, we can still do some improvements: we can add more variables to our multivariate polynomial regression, and we probably can try more groups of parameters for ARIMA model.