Linear Regression Project D208

Jacob Polomsky

June 19, 2024

A1. What factors influence outages per week ?

A2. My goal into this analysis is to is to gain better understanding of what variables correlate to outages per week.

B1. Four assumptions to linear regression, First is that the line is Linear relationship between independent variables and the dependent variable. Second making sure there is no multi-collinearity or that one independent variable does affect another independent variable. Third that normal distribution of errors. Four that each observation is independent of each other.

B2. The two benefits of Python for me is one I am more comfortable with the languages than with R. Python is a general purpose language making it more versatile than R allowing you to do a lot more with it. Python also has a extensive list of libraries that can help with running linear regression. The packages I ran in this analysis is Pandas for easy use of for the dataset. Seaborn and Matplotlib for making visualizations. Statmodels for running the linear regression model and seeing the summary statistics.

B3. The reason why I am running a multiple linear regression model is because the question being asked is “What factors influence outages per week ?” which outagesperweek in the dataset is a continuous variable. This makes it a linear relationship rather than a logistic relationship.

C1. For the Data cleaning process the first thing I did was to narrow down the dataset to use only the columns I needed for the analysis. Then I checked for duplicates to makes sure each observation is independent and found that there were no duplicates in the data set. The second thing I did was check for null values because the model cant take null values and found over 2000 null values for Internet Service and drop those them from the dataset. I used pandas .describe() function to check for outliers to makes sure the data model runs smooth and did not find any evident outliers in the data.

C2 . For the Continuous variables Population and Bandwidth have really high values compared to the other variables having a mean of around 3000. The other Continuous variable have a Age has most value fall between 35 and 70. Outages per week and Email have similar distribution being around the 10 mark. Yearly equipment failure having a really low values with most values falling within 1 or 0 and most of the values being 0. For the States it seems like a every state has drastically different counts with some having over 500 and other having less than 50, but most values falling in 300 to 100 range. Area, Martial, Internet Service, Online Backup, Device protection have a pretty even spread amongst all of their values. Tech Support has a skew more towards No, Techie has a large disparity with most being No, and gender having a even split between males and females, but Non binaries are almost non-existent.

See other word Document or D208\_Data\_Exploring.ipynb it will have a more in-depth summary for all categorical variable

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

C3. See code attached D208\_Data\_Exploring For Visualizations

C4. For Data Wrangling I had two sets of categorical variables one having two categories and another having more than two categories. If it had a yes no question, I converted yes to 1 and no to 0 with the replace function. Categorical variable that were not yes or no I used pandas .get\_dummies function to convert those values to into dummy variables.

C5 See LR\_Model\_Dataset.csv attached

D1

A screenshot of a computer

Description automatically generated

D2

I picked a backward feature elimination; this is the process helps to reduce the number of columns used in a prediction model that could lead to overfitting or throw of your analysis or just use of resources when it is not useful. What feature selection does is that it looks at the max p-value and removes that columns from the dataset because it is unneeded or is detrimental to the analysis and this function runs over and over again until the p-values is less than a significant value you pass usually 0.005. This helps to determine what rows to get rid of. I preferred this method because my data has states which mean it has a lot of dummy haves and with model metric procedures it might put to much emphasis on one metric and is harder to and determine the result of.

D3

A screenshot of a computer

Description automatically generated

E1. So, after the data reduction with backward elimination feature we were able to reduce the model by 8 original rows 11 columns in total. We found that the model is near identical to the original dataset with a R-squared of .913 for both and an adjusted R-squared of .912, and a F-statistic of under 0.00.

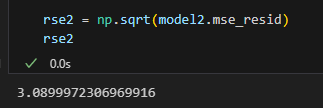
E2

A graph with a red line

Description automatically generated

A blue dotted graph with a line

Description automatically generated with medium confidence



E3. See D208\_Main\_Paper script

F1

F1A

|  |  |
| --- | --- |
| Y = 0.09(Email) + 0.23(DeviceProtection) + 6.11×10 ^−5(Bandwidth\_GB\_Year) + 8.53(State\_AL) + 8.38(State\_AR) + 8.38 (State\_AR) + 8.01(State\_AZ) + 8.07(State\_CA) + 7.90(State\_CO) + 8.36(State\_CT) + 10.02(State\_DC) + 8.32(State\_DE) + 8.31(State\_FL) + 8.53(State\_GA) + 8.10(State\_HI) + 8.39(State\_IA) + 8.86(State\_ID) + 8.17(State\_IL) + 8.25(State\_IN) + 8.25(State\_KS) + 8.01(State\_KY) + 8.05(State\_LA) + 8.29(State\_MA) + 8.17(State\_MD) + 8.34(State\_ME) + 8.07(State\_MI) + 8.34(State\_MN) + 8.30(State\_MO) +8.55(State\_MS) + 8.08(State\_MT) + 8.56(State\_NC) + 8.87(State\_ND) + 7.98(State\_NE) + 8.32(State\_NH) + 8.61(State\_NJ) + 8.23(State\_NM) + 8.45(State\_NV) + 8.32(State\_NY) + 8.07(State\_OH) + 8.32(State\_OK) + 8.45(State\_OR) + 8.13(State\_PA) + 8.16(State\_PR) + 8.19(State\_RI) + 8.39(State\_SC) + 8.60(State\_SD) + 8.28(State\_TN) + 8.41(State\_TX) + 8.08(State\_UT) + 8.28(State\_VA) +8.13(State\_VT) + 8.22(State\_WA) + 8.24(State\_WI) + 8.33(State\_WV) + 8.27(State\_WY) + 0.18(Area\_Suburban) + 0.18(Area\_Urban) + 0.19(Marital\_Married) + 0.24(InternetService\_Fiber Optic) |  |

F1B. This regression model is a bit weird it seems like depending on the state it kind of becomes the intercept the dummy variable State becomes the largest determining factor on of the equation being between 8- 10 with most values being in the 8. The rest email, device protection, bandwidth, area, marital married, and internet service all increase the number of outages by a little bit with most being around 0.15 to 0.24 and two smaller values email and bandwidth having almost no impact.

F1C . According to the summary the finding conclude that the factors are significant with a R-squared and adjusted R-squared of over .90 percent insinuates the this model explain 90% of the datapoints plotted, and also showing a f-Statistic of 0.00 showing that it is significant. But Are the finding practical in my option not really the model is not normal like the state being a weird constant variable making me delete the constant variable otherwise constant would have a high VIF and causing the model when ran to have a poor R-squared score 0.007. Makes me wonder how it would work on other data sets.

F1D. The limitations of the data, one like I mentioned above the state being a weird constant variables makes me believe the introducing more data will cause the data to stop fitting the model like it use to. I think the model will have a lot of issues when it comes to outliers throwing of the model, and this could be the possibility for the two clusters for residual plot.

F2 My recommend course of action is to keep testing it out or keep adding more data. This analysis has some hole like with residual plot having two cluster, and not having a constant variable instead the state being a weird constant variable is something I believe will cause issues going forward. But according to everything else this model does show that is a good model for predicting outcomes for Outage per week. It needs to be tested more to make sure the model doesn’t fall apart when introducing more data.