Connor Adams

Professor Armstrong

9/23/19

DS 4220-01

Assignment 2- Multiple Regression

For this assignment I had to make three regression models to predict income based off of the features that we have in our income data file, in this case it’s age, marital status, years at current address, education level, years at current employer, job satisfaction, gender, and number of people living in the home. I also had to create my own feature which I took the years at current employer and multiplied it by job satisfaction. I thought this would be a good indicator of someone’s income because if they are at their job longer and have a high-quality job satisfaction this should cause the person to have a higher income. Therefor in my feature a higher number should indicate a higher income.

The first regression model that I made was with all of the features. This includes age, marital status, years at current address, education level (ed), years at current employer (employ), job satisfaction (jobsat), gender, number of people living in the home (reside), and employ times jobsat. I wanted to see if taking all factors and putting it into a regression model would accurately depict income because it takes so many factors into account. The results I got for this were:

ME RMSE MAE MPE MAPE

Test set 3.03 66.8 34.8 -38.8 71.6

These results show that the mean error (ME) is 3.03 meaning that the average residual is 3.03 between the actual value and the predicted value. The root mean square error (RMSE) is 66.8 which shows the square difference between the actual values and the predicted values. The MAE of 34.8 is the mean absolute error which shows the average error of the absolute value of the actual value and the predicted value. The MPE of -38.8 shows mean of each residual divided by the actual outcome in percentage. The MAPE of 71.6 shows the mean of the absolute value of each residual divided by the actual outcome in percentage. These numbers are decent but we would want to get these numbers to be smaller because if we look at the ME of 3.03 this is showing the average residual was 3.03. This is not good and we would want this number to be lower. The reason in my opinion why this is so far off is because of all the features in this model that don’t matter. For example, we want to look for a p value under 0.05 to find out that the feature is significant. Since there are a lot of features that aren’t under 0.05 (gender, marital, reside) this may make the error of the model higher.

The second regression model that I made was different than the first model because I took out some of the features that I thought didn’t matter. In this case I took out age, marital, years at address (address), and reside out because I thought they wouldn’t help the data. In this model I looked at gender, job satisfaction, employ times jobsat, and education level. I thought this model would more accurately represent the data than using all of the features. The results I got were:

ME RMSE MAE MPE MAPE

Test set 2.99 66.9 34.7 -39.4 71.6

These results show that the mean error (ME) is 2.99 meaning that the average residual is 2.99 between the actual value and the predicted value. The root mean square error (RMSE) is 66.9. The MAE of 34.7. The MPE of -39.4 which shows mean of each residual divided by the actual outcome in percentage. The MAPE of 71.6. These numbers ended up turning out to be not as good as I thought they were going to be. I thought that taking out some of the features I thought were insignificant would help. It did in a little way with the mean error being 2.99 which is pretty good but the RMSE, MAE, MPE, and MAPE were all still around the same as the original model, which to me seems to be too high. The only feature in this model that wasn’t significant was gender at 0.09.

The third regression model I made was different than both the original regression model and the second regression model. I took out some of the features that I thought may have had a significant effect on the model but were covered in my new feature of employ times job sat, because I felt that there may be no need to include the employ and job sat in the model. I also took out ed (education level), to see if this would make the model any better to just have gender, age, marital, reside, address, employtimesjobsat. The reason I did this was to create a model that was different than the original and the other model I created. The results I got were:

ME RMSE MAE MPE MAPE

Test set 3.93 70.8 36.8 -46.5 71.6

These results show that the mean error (ME) is 3.93 meaning that the average residual is 3.93 between the actual value and the predicted value. The root mean square error (RMSE) is 70.8. The MAE of 36.8. The MPE of -46.5 which shows mean of each residual divided by the actual outcome in percentage. The MAPE of 71.6. These numbers turned out to be somewhat what I thought with the ME being so high because I knew that taking out features that were significant would make the data look like this. This model is the worst one because all of the values are higher than the other models. The only feature in this model that is significant is the employtimesjobsat. The value that seems most alarming to me is the MPE of -46.5 because this shows that the mean percent error is 46.5% which is significantly higher than the other two models. This model to me seems like a poor model and doesn’t relate well to finding income based off of the features we had.

In conclusion the model that I would recommend to find income based off of the features we were given is the second model. The reason why I say this is because of the ME number. I thought that the mean error was a significant value in determining which model to recommend because it shows the mean error between the actual value and the predicted value. If you look at all of the models that I made, model 2 has the lowest ME. This shows that it is the most accurate regression out of all three with the data set given. We can conclude that the third model is the worst since all of it’s numbers were so high and we want smaller numbers. However, the decision between using the first model and the second model was close. I thought that all of the numbers between RMSE, MAE, MPE, and MAPE were all very close between the two. Since they were so close, I thought the deciding factor was the ME and since the ME was smaller in model 2 that is the one that I would recommend. Also, with model 2 there are less features in that model so the data doesn’t get clustered and the percent of significant features in the model is higher.

I would recommend using REGRESSION MODEL 2.