4. Modelling

The base model consisted of all the features available in the dataset. This model is a Linear-Model with no interaction between the features. The included features in this model were: ***Population, BMI, Year, Total Expenditure (on healthcare), Continent, Polio, Diphtheria, Status,*** and ***Schooling***. The purpose of this model was to get a picture of how all the features affected Life Expectancy, or whether they affected the Life Expectancy at all. The features with lower values of co-efficients were eliminated from further analysis. As these features have little to no effect on the Life Expectancy. The features eliminated, on the basis of their co-efficients in the model, were ***Population, Year***. Another feature was eliminated i.e. ***BMI*** on the basis of having a large amount of null values, which affected the overall predictive power of the model. Population reason ? The reason the Year is not such a good predictor could be that, even though the Life Expectancy is increasing as the years go by, but there are a lot of underlying factors at play. Hence there is much more to the story rather than just Life Expectancy simple increasing by the years. Diphtheria was also removed from the final model as it is highly correlated with Polio (r=0.7).

After eliminating the above-mentioned features, different combinations of models with interaction between different features was tried. Based on the lowest AIC score amongst these models, the model selected was with predictors: ***Polio, Total Expenditure, Schooling, Continent, Continent \* Total Expenditure, Continent \* Schooling*** … (*where x1 \* x2 signifies interaction between features x1 and x2*). The r-squared score for this model was 0.75, which indicates that this model is a good enough model for the purpose of predicting Life Expectancy. The included interactions are essential as they are self-explanatory. ***Continent*** provides us with a generalized sense of how other factors affect Life Expectancy and how they vary geographically. This is extremely important to our analysis as earlier we saw that the ***Total Expenditure*** varied by each ***Continent***, and the effect of ***Total Expenditure*** on ***Life Expenditure*** was different for each ***Continent.*** Hence the interaction term between them bolsters the model predictive power. Apart from this Schooling is a very important feature as it is very highly correlated with ***Life Expectancy***. ***Schooling*** again varies by each ***Continent***; this is mostly based on the fact if the country is Developing or Developed. As we saw for countries with higher Schooling, the Life Expectancy was higher as well, and these were mainly Developed nations, however some of the Developing countries were a part of this bracket as well. This further affirms the strong effect of the years of ***Schooling*** on the overall ***Life Expectancy.*** And ***Schooling*** was different for countries based in different continents, for instance, there was a divide in the years of education if we compare countries in North America to those in Africa.

But to further explore how varied of an effect did ***Schooling***, as well as ***Total*** ***Expenditure***, had on ***Life Expenditure*** we decided to generate two models trained with the same structure but only on the subset of data of years 2000 and 2014. This is because, 2000 to 2014 would be a large enough time-period to notice differences between ***Life Expectancy*** and all the factors affecting ***Life Expectancy***. And comparing the results from both these models would help us have a contrastive comparison on how things changed between these years.

After training the models for years 2000 and 2014, the goal was to check the trends for the fitted values for the two models. Dummy data was simulated for both years, by generating combinations of values for the predictors. Using these simulated values, the values were fed to the model to predict the **Life Expectancy**. Using these predicted values, the following plots were used to compare the results of the two models for each of the metrics (i.e., ***Schooling, Total*** ***Expenditure*** & ***Polio)***.

1. Schooling:

Chart, line chart

Description automatically generated

Figure 4.1: Schooling and predicted Life Expectancy

In Figure 4.1, we can observe that in both the years 2000 and 2014, there is a strong positive correlation between Schooling and Life Expectancy, for each continent. But for the year 2000, there is a lot more variation between the continents as compared to year 2014. As for Africa the trend has just shifted upwards from 2000 to 2014. One reason explaining this variation could be that, as for the recent years, the emphasis on education has increased as compared to the earlier years. This impact of more emphasis on education makes the feature ***Schooling*** a much stronger and direct contributor to a better ***Life Expectancy*** value for every country, for the year 2014. This explains the uniformity that can be noticed for every continent for that year. In other words, we can say the for the recent years, Schooling is directly, strongly and positively correlated to Life Expectancy, and this is the case with every continent.

1. Polio

Chart

Description automatically generated

Figure 4.2: Polio and predicted Life Expectancy

In Figure 4.2, we can see that the ***Polio*** vaccination rate has always had a positive correlation with life expectancy. Also, the overall ***Life Expectancy*** for Africa has gone up by 10 years. The life expectancy of Asia also increased by nearly 5 years. We can say that immunizations always have had a strong effect on ***Life Expectancy***.

1. Total Expenditure:

Chart, line chart

Description automatically generated

Figure 4.3: Total Expenditure and predicted Life Expectancy

In Figure 4.3, for the year 2000, North America has a slight downwards ***Total Expenditure*** trend, as opposed to other continents, except Africa. However, the trend for North America is quite flat. As for Asia, the trend goes from flat to slightly increasing from the years 2000 to 2014. Africa, as we saw before, is a special case here. The trend is downwards for both the years 2000 and 2014. But it can be observed that for the year 2000, Africa has a much steeper trend as compared to the year 2014, signifying the increase in Life Expectancy in Africa, amongst the two years 2000 and 2014. However, the trend is still interestingly downwards. This can be attributed to our hypothesis earlier in the EDA section, where even with the higher amount of healthcare expenditure, the prevalence of HIV/AIDS affects Life Expectancy inversely.