Final Report

Machine Learning

Project Title:

Prediction/Analysis of Healthy Life Expectancy

Prepared By:

Simul Bista (N01489966)

Kaiyan Chen (N01489178)

ITE 5332- Winter 2023
Humber College

Problem Statement

This project aims to use machine learning models to predict healthy life expectancy and identify key factors influencing it, based on data from the World Happiness Report (2005-2022). Healthy life expectancy is a measure of the expected number of years a person can live in good health.

Dataset Description

- The dataset is taken from the World Happiness Report (2005-2022) based on Gallup World Poll data.
- The following features were trained to predict/analyze the healthy life expectancy (years) out of many:
  - Happiness Score (measure of life satisfaction (0 - 10, 10 = best possible life, 0 = worst possible life)
  - GDP per capita
  - Social Support (national average of responses about social support. (0-1, 1 = yes, 0 = no)

Dataset Analysis & Observations

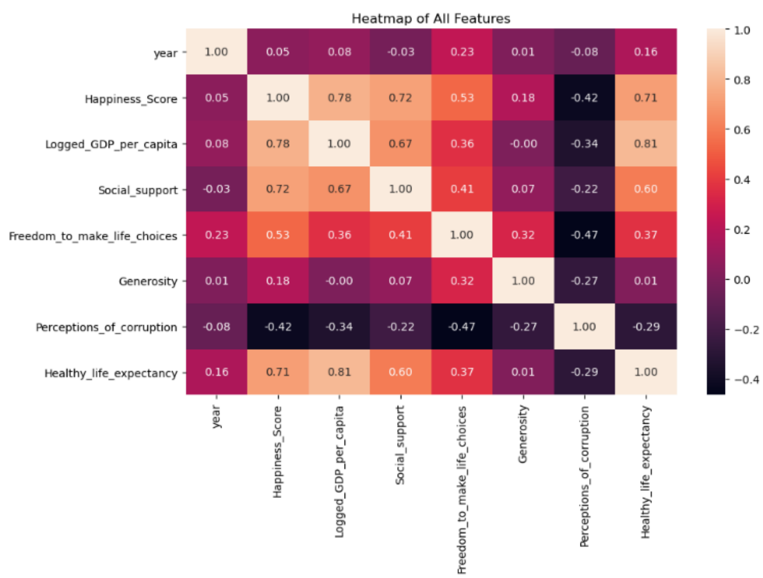

Fig 1: Heatmap (correlation) between the features

- Initially the data was cleaned (null values were replaced with the mean)
- Univariate and Bivariate analysis were carried out to understand the correlation between the features and the target variable i.e., the healthy life expectancy.
- The heat map (Fig. 1) shows that the features- **happiness_score, gdp_per_capita** and **social_support** has **strong correlation** with healthy_life_expectancy with correlation coefficient of 0.71, 0.81 and 0.60 respectively.
- The pair plot (Fig.2) shows that healthy_life_expectancy has **moderate, non-linear, and positive correlation** with the three features.
- Finally, the univariate analysis of healthy_life_expectancy (Fig. 3) shows **negative skewness** which indicates the need for non-linear algorithms).

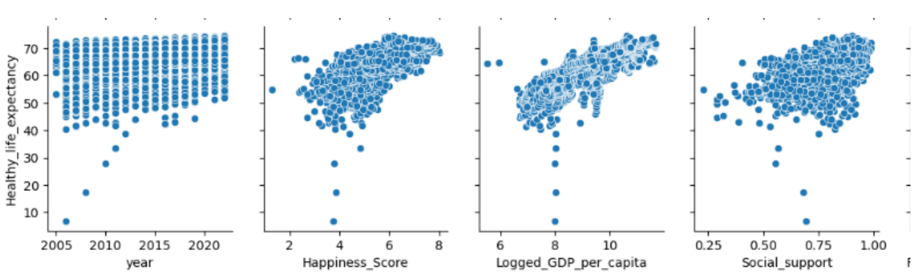

Fig 2: Pair plot between healthy life expectancy and other features

Fig 3: Univariate Analysis (price)

Proposed Analytical/Prediction Model

- There were 3 dependent and continuous variables namely Happiness_Score, Logged_GDP_per_capita and Social_support that showed correlation to healthy life expectancy, hence polynomial regression and feed forward neural network were preferred.
- For Polynomial Regression Model, degree 3 was used which swayed away from overfitting/underfitting behavior and hence better result (with less errors) was achieved.
- For Feedforward Neural Network, 3 hidden layers with 30 neurons were used. Increasing the number of neurons in each hidden layer along with increasing the number of hidden layers in the modeling – resulted in a better model with reduction of training loss and validation loss. Increasing the neurons and the number of layers further did not change the result significantly.

Results & Discussions

- The **polynomial model (with degree 3)** performed the best with the **least errors** (in fig 4).
- **Hyper parameter tuning** improved the model as seen in fig 5.
- The model performance comparison as seen in fig 6 showed slightly better performance by the regression model.
- **MSE comparison** for both models in fig 7 proved regression to be the better performer for the dataset.

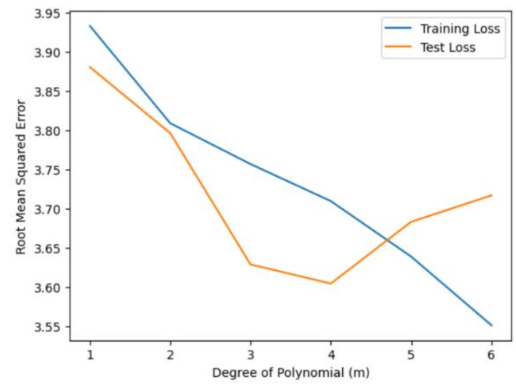

Fig 4: Determining the best value of degree for Regression Model

```
# Define the forward neural network
ffnn_model = Sequential()
ffnn_model.add(Dense(20, input_dim=3, activation='relu'))
ffnn_model.add(Dense(30, activation='relu'))
ffnn_model.add(Dense(30, activation='relu'))
ffnn_model.add(Dense(30, activation='relu'))
ffnn_model.add(Dense(1, activation='linear'))

# Compile the model
ffnn_model.compile(loss='mean_squared_error', optimizer='adam')

# Fit the model
ffnn_model.fit(X_train, y_train, epochs=10, validation_split=0.1)
predictions_ffnn = ffnn_model.predict(X_test)

Epoch 1/10
50/50 [==============================] - 1s 5ms/step - loss: 3994.5579 - val_loss: 3892.5349
Epoch 2/10
50/50 [==============================] - 0s 2ms/step - loss: 2933.8162 - val_loss: 1146.9879
Epoch 3/10
50/50 [==============================] - 0s 2ms/step - loss: 608.6876 - val_loss: 340.6044
Epoch 4/10
50/50 [==============================] - 0s 2ms/step - loss: 336.3173 - val_loss: 197.9286
Epoch 5/10
50/50 [==============================] - 0s 2ms/step - loss: 197.2556 - val_loss: 111.9390
Epoch 6/10
50/50 [==============================] - 0s 2ms/step - loss: 110.9300 - val_loss: 64.1891
Epoch 7/10
50/50 [==============================] - 0s 2ms/step - loss: 68.0295 - val_loss: 42.3304
Epoch 8/10
50/50 [==============================] - 0s 2ms/step - loss: 47.0625 - val_loss: 30.4949
Epoch 9/10
50/50 [==============================] - 0s 2ms/step - loss: 35.2971 - val_loss: 23.9328
Epoch 10/10
50/50 [==============================] - 0s 2ms/step - loss: 28.2816 - val_loss: 20.2904
14/14 [==============================] - 0s 924us/step
```

Fig 5: Hyperparameter Tuning in Feed Forward Neural Network

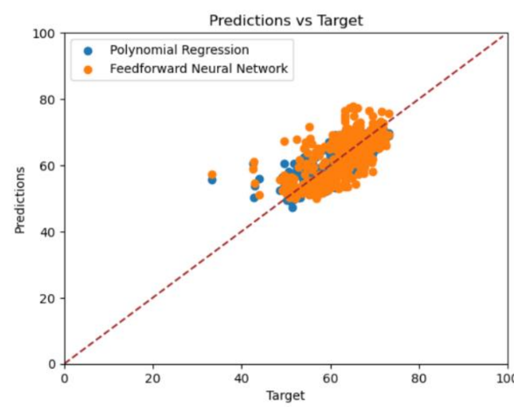

Fig 6: Predictions Vs Targets for both models

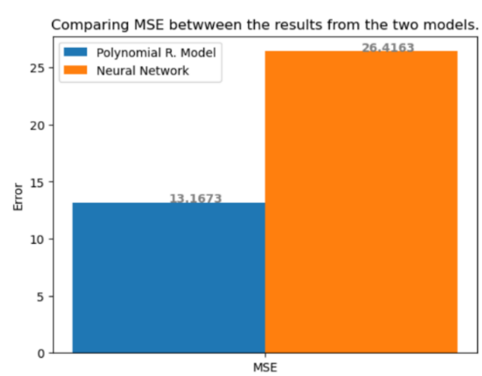

Fig 7: MSE comparison between both models

Conclusion

The analysis concludes that features such as **GDP per capita**, **social support**, and the **happiness degree** impact the healthy life expectancy compared to other features. Moreover, improved data accuracy was observed in the **polynomial regression model** (**of degree 3**) in comparison with feedforward neural network, to predict the healthy life expectancy. This is because of the small size and low complexity of the dataset.