Life Expectancy: A Data-Driven Exploration and Prediction

Milestone: Model Exploration and Model Selection

Group 13

Sai Varun Kumar Namburi

Prajwal Srinivas

namburi.sai@northeastern.edu

srinivas.pra@northeastern.edu

Percentage of Effort Contributed by Student1: 50%

Percentage of Effort Contributed by Student2: 50%

Signature of Student 1: Sai Varun

Signature of Student 2: Prajwal

Date of Submission: 20th Mar 2023

Data Source: This dataset is taken from kaggle.com
https://www.kaggle.com/code/philbowman212/life-expectancy-exploratory-data-analysis/data

Data Description:

The World Health Organization's Global Health Observatory (GHO) maintains records of the health status and related factors of all countries. The data concerning life expectancy and health factors for 193 countries were obtained from the WHO's Global Health Observatory data repository. It was noted that over the past 15 years, there has been significant progress in the health sector, leading to a significant improvement in human mortality rates, particularly in developing nations compared to the last 30 years. In this project, data from the years 2000 to 2015 for 193 countries were selected for further analysis. In this dataset, we have 22 columns as below
(Country, Year, Status, Life expectancy, Adult Mortality, infant deaths, Alcohol, percentage expenditure, Hepatitis B, Measles, BMI, under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness 1-19 years, thinness 5-9 years, Income composition of resources, Schooling)

Feature Selection:

The life expectancy dataset contains a number of features that could potentially be used to predict life expectancy. Here are some possible steps for feature selection.

1. Identify the target variable: In this case, the target variable is life expectancy.
2. Explore the data: Look at the distribution and relationships between each feature and the target variable. This can be done using descriptive statistics, visualizations (e.g. scatterplots, histograms, boxplots), and correlation matrices.
3. Consider feature relevance: Determine which features are likely to be relevant for predicting life expectancy based on prior knowledge, domain expertise, and common sense. For example, factors such as income, access to healthcare, and education are likely to be relevant.
4. Remove redundant features: If there are multiple features that are highly correlated with each other, consider removing one of them to avoid overfitting.
5. Use feature selection techniques: There are a variety of feature selection techniques that can be used to select the most important features for predicting life expectancy.
   These include
   - Correlation-based feature selection: This method involves selecting features that are highly correlated with the target variable.
   - Wrapper methods: These methods involve selecting subsets of features and evaluating the performance of a predictive model using only those features.
   - Embedded methods: These methods involve incorporating feature selection into the process of building a predictive model, such as by using regularization techniques like Lasso or Ridge regression.
6. Evaluate the selected features: Once a set of features has been selected, evaluate the performance of a predictive model using only those features. This can be done using techniques such as cross-validation or hold-out validation

7. Iterate: If the performance of the predictive model is not satisfactory, consider revisiting previous steps and making adjustments to the feature selection process.

Ultimately, the choice of feature selection method will depend on the specific characteristics of the dataset and the goals of the analysis.

FEATURE SELECTION

+ Code + Markdown

```
import scipy.stats as stats
stats.ttest_ind(data.loc[data['Status']=='Developed','Life_Expectancy'],data.loc[data['Status']=='Developing','Life_Expectancy'])
```

[52]

```
Ttest_indResult(statistic=29.36614038467105, pvalue=6.43229833545797e-166)
```

p value is < 0.05 Therefore, the difference of Life Expectancy between Developed and Developing countries is significant. We can consider 'Status' as a feature.

Also "Adult mortality" : -0.7, "HIV/AIDS" : -0.56, "BMI" : 0.57, "Polio" : 0.47, "GDP" : 0.46, "Alcohol" : 0.4, "thinness_1to19_years" : 0.45

Feature engineering is the process of selecting, transforming, and creating new features from the raw data to improve the performance of machine learning models. The goal of feature engineering is to make the data more informative and easier for machine learning algorithms to process.

In the context of life expectancy data, feature engineering can involve creating new variables or transforming existing ones to better capture the factors that influence life expectancy. Some examples of feature engineering for life expectancy data include:

FEATURE ENGINEERING

- **DUMMIFICATION OF STATUS**
- **NORMALIZING NUMERICAL FEATURES**

```
feature_df = data[['Country','Status','Adult_Mortality','Alcohol','HIV/AIDS','Polio','BMI', 'thinness_1to19_years','Life_Expectancy']]
```

[53]

```
feature_df = pd.concat([feature_df,pd.get_dummies(feature_df['Status'],drop_first=True)],axis=1)
final = feature_df.drop('Status',axis=1)
```

[54]

1. Creating aggregate measures: Aggregating data for each region by taking the mean or median of several variables such as education, income, and healthcare access can help capture the overall quality of life in each region.
2. Scaling and normalization: Scaling and normalization of the data can make features more comparable and easier for models to work with. This is particularly useful for variables with different units of measurement, such as income and education.
3. Time-series features: Creating new variables that capture trends, seasonality or periodicity in life expectancy data over time can be used to help models account for the temporal dynamics in life expectancy.

4. Feature interaction: Interactions between features can help to better capture their joint effects on life expectancy. For instance, interaction features such as the product of education and income could provide a better measure of socioeconomic status than each variable individually.
5. Age transformation: Age is one of the most important predictors of life expectancy, and it's relationship with life expectancy is often nonlinear. Transforming age by taking its square or logarithm can help capture this nonlinearity.

```
[56]
final['Polio_scaled'] =final['Polio'].apply(lambda x : ((x-np.min(final['Polio']))/(np.max(final['Polio'])-np.min(final['Polio'])) * (20)))

[57]
final.to_csv('./final.csv',index = False)

[58]
final = pd.read_csv('./final.csv')

[59]
final.head()
```

	Country	Adult_Mortality	Alcohol	HIV/AIDS	Polio	BMI	thinness_1to19_years	Life_Expectancy	Developing	Adult_Mortality_scaled	Polio_scaled
0	Afghanistan	263.0	0.01	0.1	6.0	19.1	17.2	65.0	1	7.257618	0.625000
1	Afghanistan	271.0	0.01	0.1	58.0	18.6	17.5	59.9	1	7.479224	11.458333
2	Afghanistan	268.0	0.01	0.1	62.0	18.1	17.7	59.9	1	7.396122	12.291667
3	Afghanistan	272.0	0.01	0.1	67.0	17.6	17.9	59.5	1	7.506925	13.333333
4	Afghanistan	275.0	0.01	0.1	68.0	17.2	18.2	59.2	1	7.590028	13.541667

Feature engineering can help to identify and capture the most important factors that influence life expectancy. This, in turn, can improve the performance of machine learning models, providing more accurate predictions of life expectancy. By carefully selecting and transforming features, feature engineering can help to create more powerful models that can better understand and predict life expectancy.

EMBED THE COUNTRY FEATURE

+ Code + Markdown

```
[60]
countries = final.Country.unique()
country_dict = {'countries': list(countries)}
country_df = pd.DataFrame(country_dict)

[61]
def demo(feature_column):
    feature_layer = layers.DenseFeatures(feature_column)
    return feature_layer(country_dict).numpy()

[62]
countries = feature_column.categorical_column_with_vocabulary_list(
 'countries', country_df['countries'])

[63]
countries_embedding = feature_column.embedding_column(countries, dimension=4)

[64]
countries_embedding =  demo(countries_embedding)

[65]
b = []
for embed in countries_embedding:
    b.extend([embed] * 16)
```

```
[66]  final['countries_embedding'] = pd.Series(b)

[67]  final['sum_countries_embedding'] = final['countries_embedding'].apply(lambda x: x.sum())

[68]  final =  final.rename(columns={"HIV/AIDS": "hivaids"})

[69]  feature_df = ['sum_countries_embedding','Adult_Mortality_scaled','Alcohol','hivaids','Polio_scaled','BMI', 'thinness_1to19_years','Developing']
```

Model Exploration:

Model exploration involves understanding the performance and behavior of a model on a given dataset. Here are some steps for model exploration with the life expectancy data:

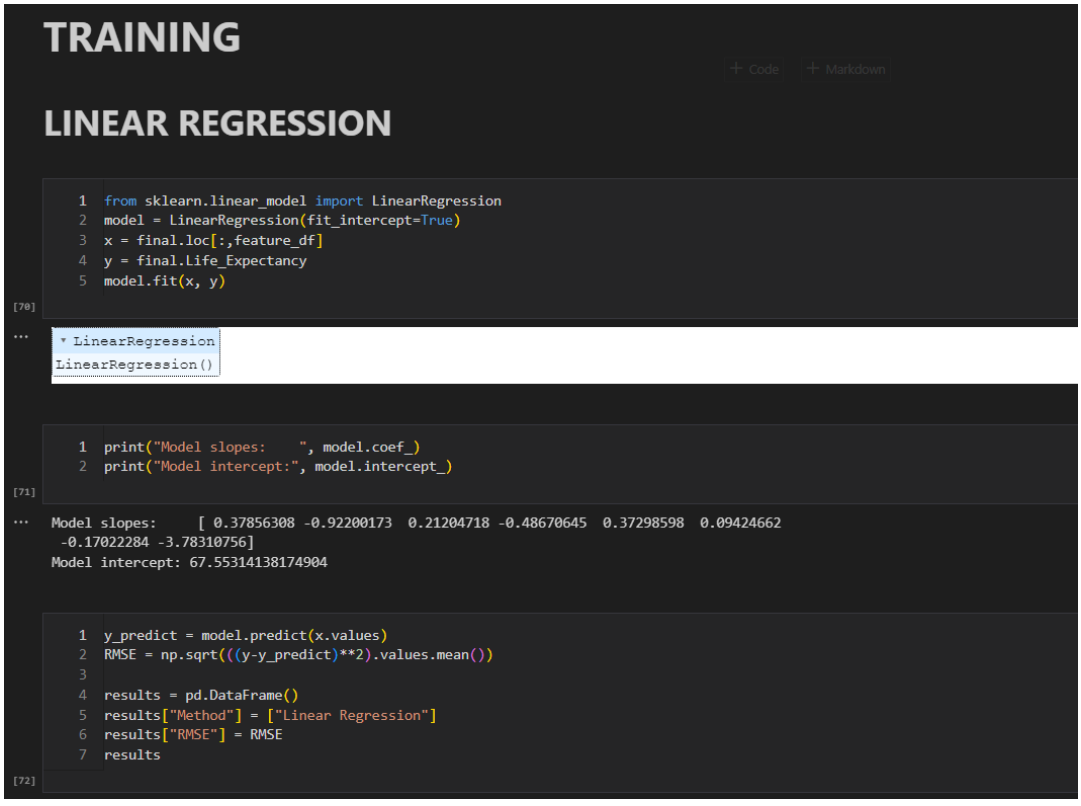
1. Split the data into training and testing sets: Divide the dataset into two parts - a training set and a testing set. The training set will be used to train the model, and the testing set will be used to evaluate the model's performance.
2. Train a baseline model: Start by training a simple baseline model, such as linear regression, to establish a baseline level of performance.
3. Explore hyperparameters: Explore the hyperparameters of the model to find the best combination for the given data. For example, in a decision tree model, hyperparameters may include the maximum depth of the tree, the minimum number of samples required to split a node, and the maximum number of features to consider at each split.
4. Evaluate performance: Evaluate the performance of the model on the testing set using appropriate metrics, such as mean squared error or R-squared for regression models, or accuracy and precision for classification models.
5. Visualize results: Visualize the results of the model exploration process to gain insights into the behavior of the model. For example, plot the learning curves to see how the performance of the model changes as the size of the training set increases.
6. Repeat: Iterate through the steps above, trying out different models and hyperparameters to improve the performance of the model.
7. Explain the model: Finally, it's important to understand how the model is making predictions. For example, for a decision tree model, you can visualize the decision rules used to make predictions. For more complex models like neural networks, you may need to use techniques like LIME or SHAP to understand how the model is making decisions.

By exploring different models and hyperparameters, and visualizing the results, you can gain insights into the behavior of the model and make informed decisions about which model to use for predicting life expectancy.

Different Types of Models possible for the Life expectancy Dataset

There is no "best" model for the life expectancy data, as the choice of model will depend on the specific characteristics of the dataset and the goals of the analysis. However, some commonly used models for predicting life expectancy include linear regression, decision trees, random forests, and neural networks.

1. Linear regression is a simple and interpretable model that can be used to identify the relationship between life expectancy and a set of predictor variables. However, it assumes a linear relationship between the target variable and the predictors and may not capture more complex relationships.
2. Decision trees and random forests are tree-based models that can capture nonlinear relationships and interactions between features. They are also relatively interpretable, as the decision rules used to make predictions can be visualized.
3. Neural networks are a powerful class of models that can capture complex relationships and interactions between features. However, they can be difficult to interpret and may require more data and computational resources than simpler models.
4. Ultimately, the choice of model will depend on the goals of the analysis, the size and quality of the dataset, the computational resources available, and the trade-off between model interpretability and performance. It is important to evaluate the performance of multiple models and compare their strengths and weaknesses before making a final decision.



Here we are training the data using the Linear Regression Model for the life expectancy dataset.

In the future, we are going to explore the Decision Tree, Random Forest, Mixed Effect Model, and Neural Networks Model
Then we can compare the accuracy of each and every model and we can easily identify which model is a good fit for this data.