Predicting Happiness using Gallup World Happiness Data

Cynthia Hester
Data Science Institute, Brown University
https://github.com/HestC/data1030_project.git

1 Introduction

The Gallup World Happiness Report, spanning 2005 to 2022, provides a comprehensive assessment of global well-being by exploring economic, social, and environmental influences. It reveals consistent happiness patterns among 165 nations, highlighting the impact of geopolitical and economic changes. Key factors integral to the report's analysis include income, social support, life expectancy, freedom of choice, generosity, and perceptions of corruption. Serving as a practical tool for policymakers and researchers, it offers insights into elements contributing to a satisfying life and guides efforts to foster happiness. The annual report ranks countries based on happiness levels using the Cantril ladder, a scale from 0 to 10 where respondents rate their own lives. With surveys conducted annually, the Gallup World Poll spans nationally representative samples from 2005 to 2022, involving over 100,000 people in 165 countries each year, providing a snapshot of global happiness trends.

About the data

The dataset was sourced from Kaggle[1] and is based on survey data from the Gallup World Happiness Poll.The World Happiness Poll bases its happiness scores and rankings on data from the Gallup World Poll which started in 2005, this poll continuously surveys citizens from 165 countries, accounting for more than 98% of the global adult population. The poll includes over 100 global questions along with region-specific ones. It comprises 2199 rows and 13 features. It has 2 categorical features and 11 continuous. It is not independent and identically distributed (non-iid), due to group structure and temporal features. Because it is non-iid, it is deterministic and thus has no standard deviation. Due to it being continuous,it is a regression problem.

Motivation

The goal of this project is to utilize machine learning techniques to analyze the **Gallup World Happiness Polls** from 2005 to 2022, specifically focusing on the **Happiness**

Score or Life Ladder Score as the target variable. By training these models on historical data,this predictive capability has the potential to assist policymakers in identifying key factors influencing happiness and implement targeted interventions to address specific challenges faced by global populations.

2 Exploratory Data Analysis (EDA)

I used the .describe method to inspect the dataset for missing values and determine the minimum and maximum values. A histogram distribution plot, to visualize the distribution of the renamed target variable 'Happiness_Score' and relationships is used.

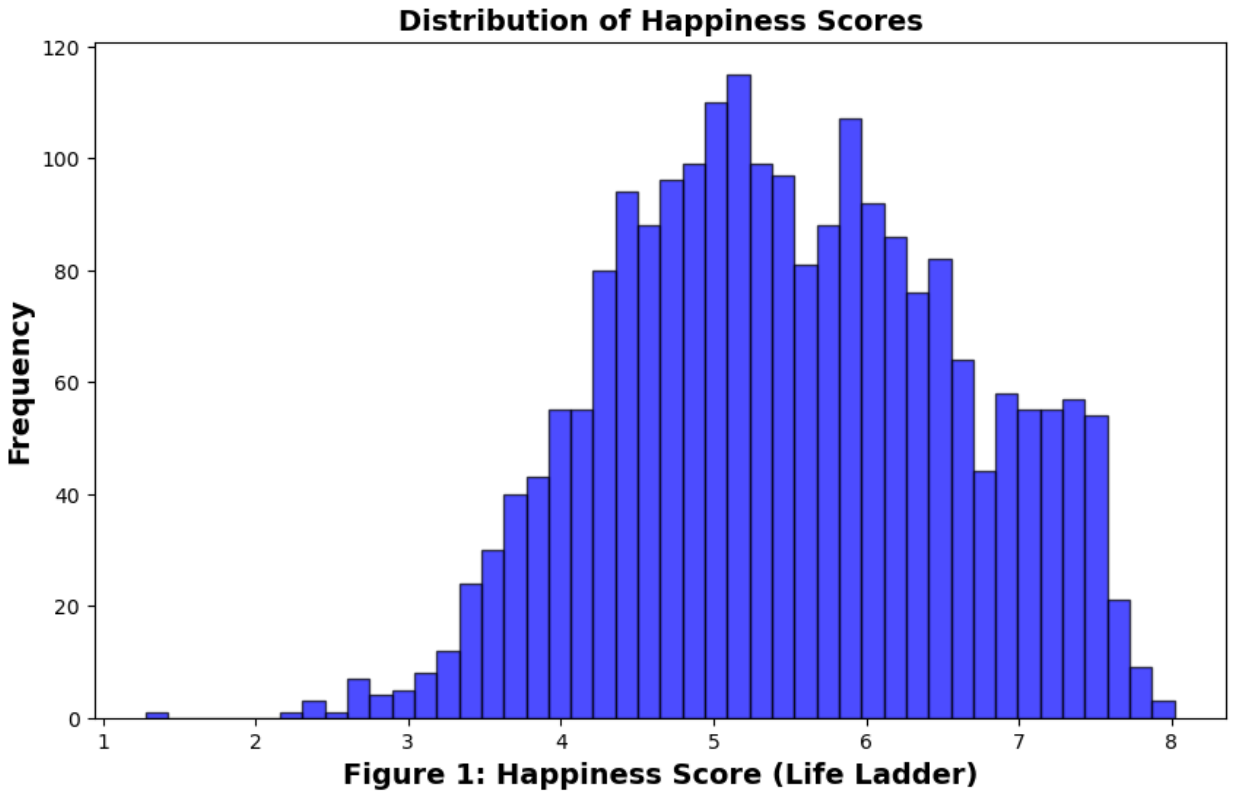


Figure 1: Happiness Score (Life Ladder)

Histogram of the distribution of the target variable 'Happiness Score' where the x-axis is the **happiness score**, and the y-axis is the **frequency.** The plot shows that the most common happiness score is around 6, and that the distribution is approximately normal.

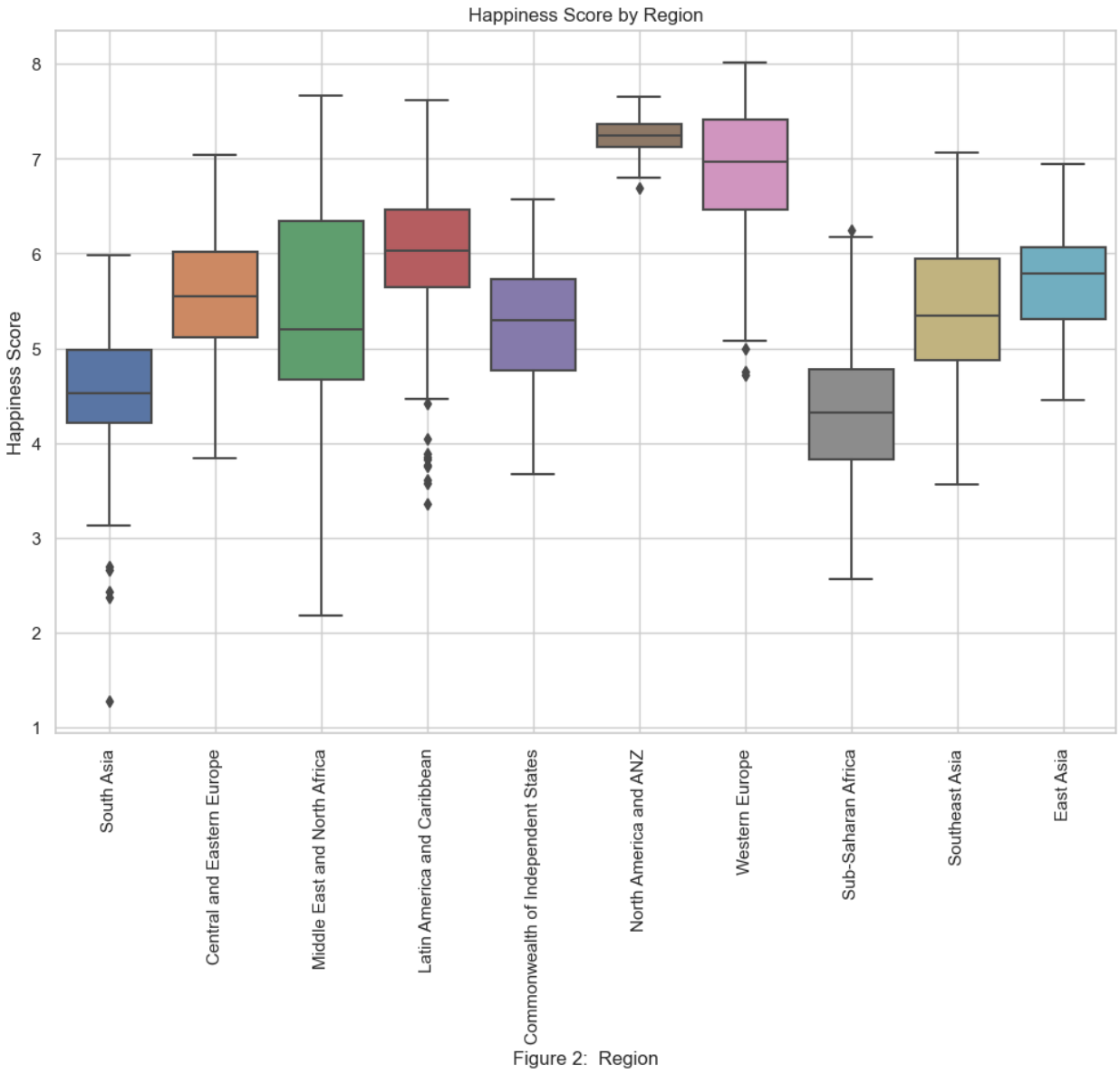


Figure 2: Region

Figure:2 The x-axis is labeled “Region” and the y-axis is labeled “Happiness Score”. The graph shows that the happiness score is generally higher in the Americas and Europe and lower in Asia and Africa.

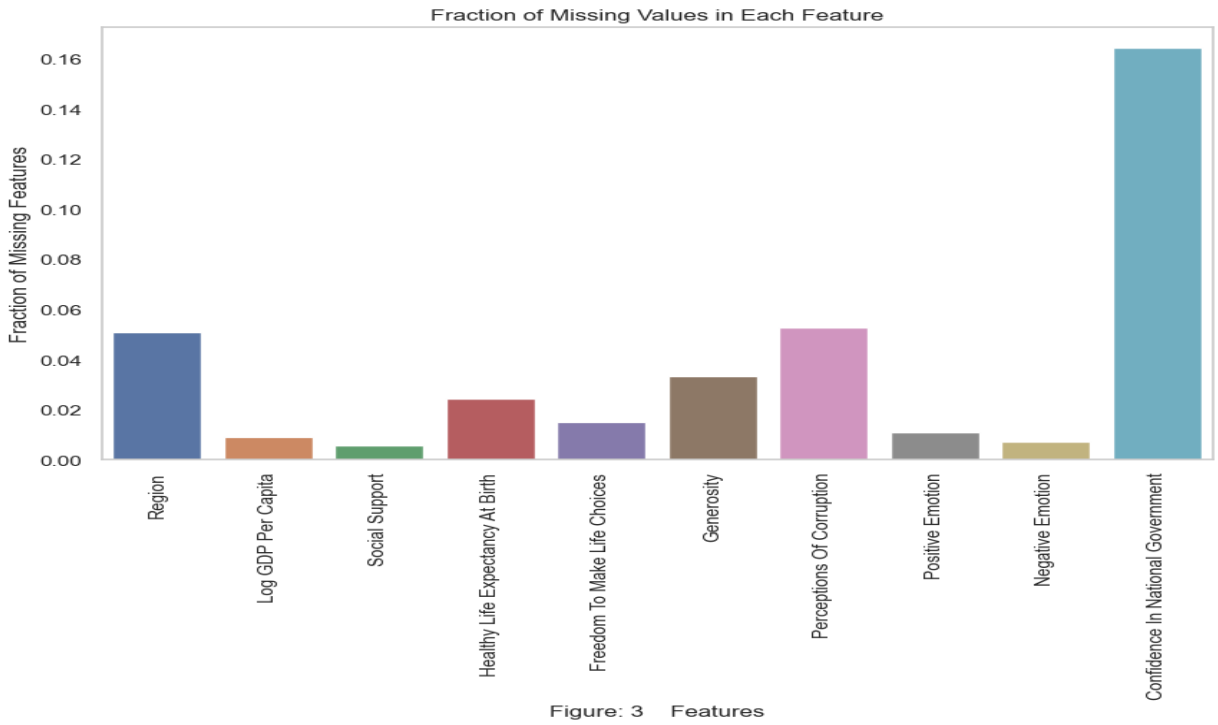
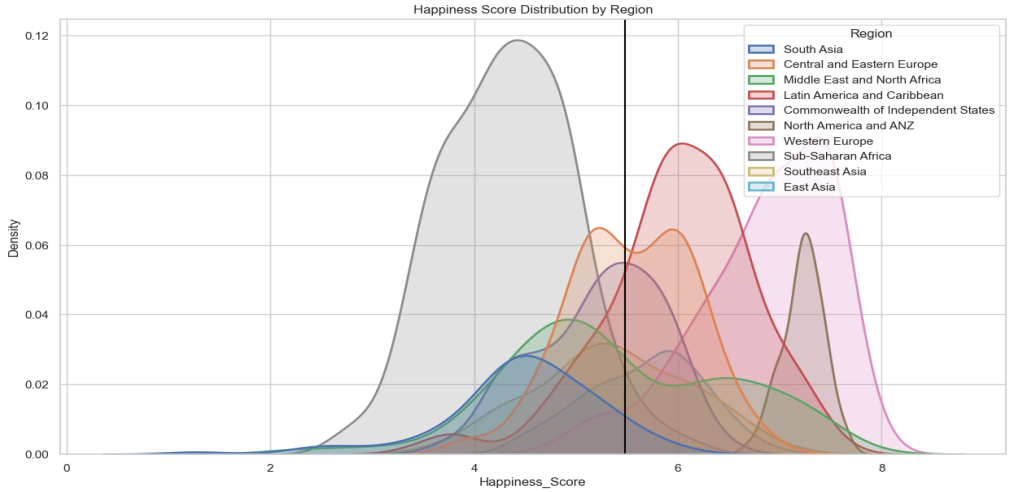


Figure 3: 10 out of the11 numeric features contain missing values.



Density plot of happiness scores by region. The plot shows that happiness scores are highest in North America and ANZ, and lowest in South Asia.

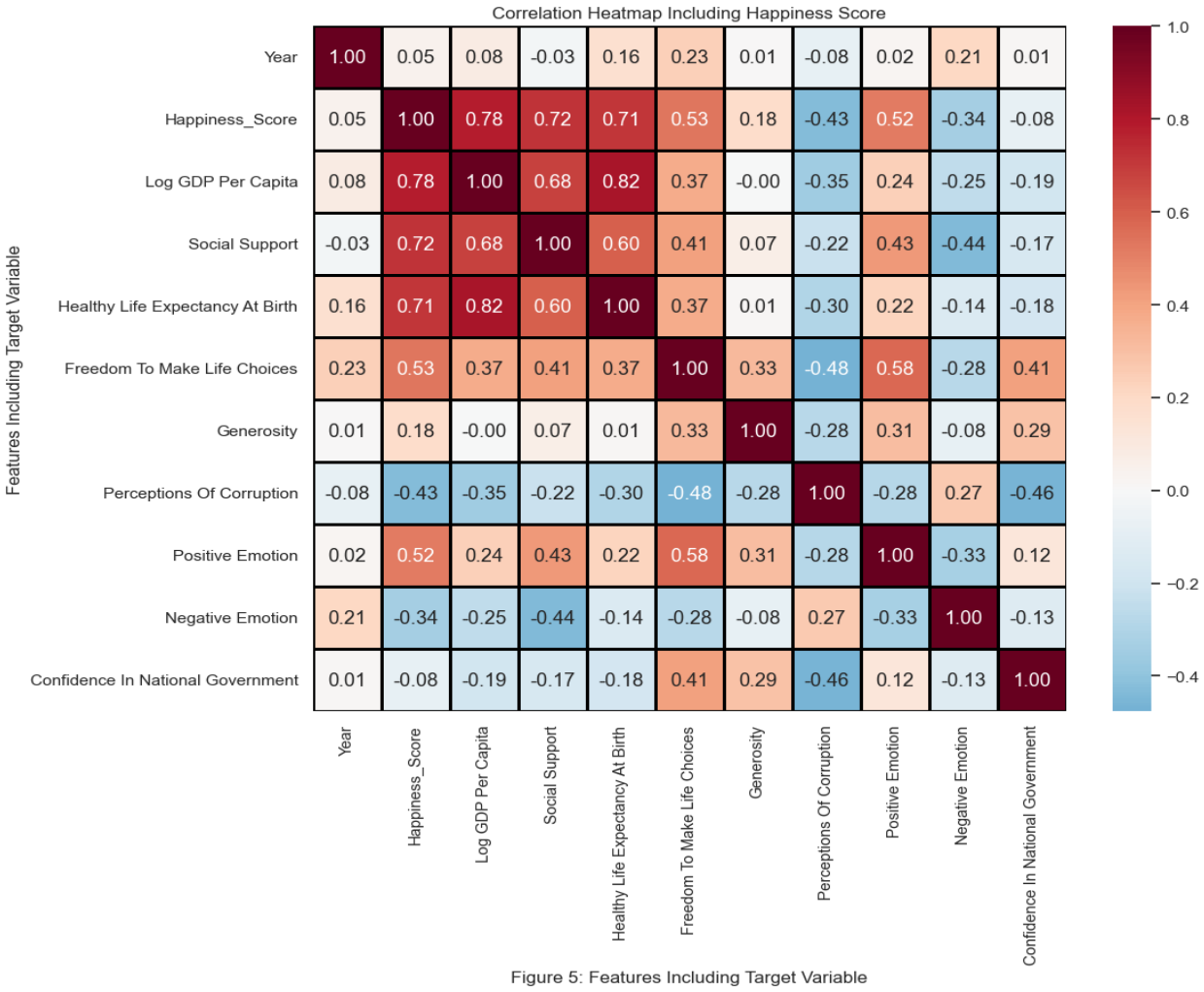


Figure 5: Features Including Target Variable

Figure 5: The heatmap shows that there is a strong positive correlation between **'Happiness Scores' and ' LogGDP per Capita', 'Healthy life Expectancy'**,indicating that countries with higher GDP and better health tend to have higher happiness scores. On the other hand, it is noticed that there is a negative correlation between the "Happiness Score" and the "Perceptions of Corruption" variable, implying that higher levels of perceived corruption are associated with lower happiness scores.

3 Methods

The dataset, characterized by its non-IID nature due to group structures (countries) and the temporal element (year), necessitated a specialized approach for data splitting. This was crucial to ensure both the integrity of the time series within each

group and the validity of the subsequent analysis. The data was initially grouped by country. This step acknowledges the distinct patterns and characteristics inherent to each country's data, treating each as a separate time series entity. Thus, the Time-Series Split package in scikit-learn, which is a cross-validator tailored for time series data, was used for sequential train-test splits for improved model evaluation.
For preprocessing the data, the pipeline started with the categorical data using the OneHotEncoder. This step was necessary to convert categorical variables into a numerical format. The OneHotEncoder facilitated the creation of binary columns, ensuring compatibility with the algorithms. StandardScaler was then applied after OneHotEncoding to normalize coefficients for later analysis of feature importance. The categorical features converted were **'country name**' and **'region'.** After these features were converted, data dimensionality increased from 12 to 174.

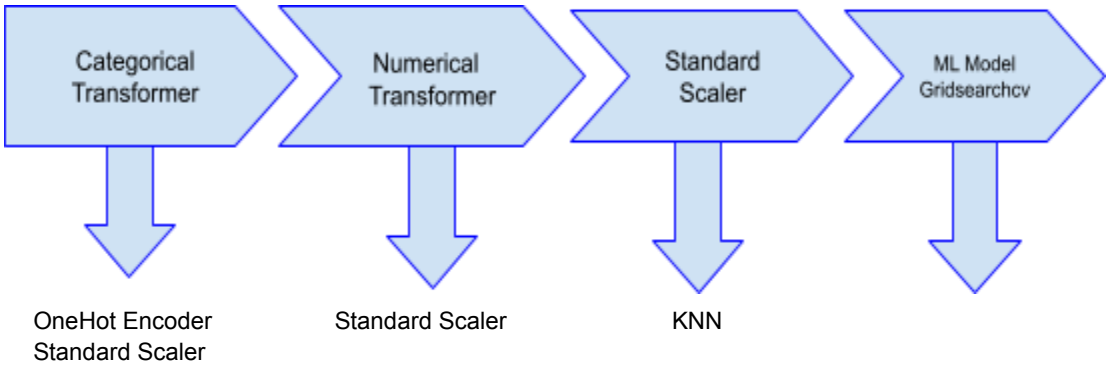


Figure:6 Pipeline for preprocessing categorical and continuous data

The mean happiness score is used with autocorrelation analysis. This helps simplify the temporal dynamics by aggregating data at the regional level, making it easier to identify overall trends and patterns for global regions. Since the dataset contains happiness scores for multiple countries within global regions for each year, taking the average allows for a consolidated view that helps clarify individual regions variabilities and outliers, providing insight into the temporal dependencies and shifts in the happiness levels of the region over time.

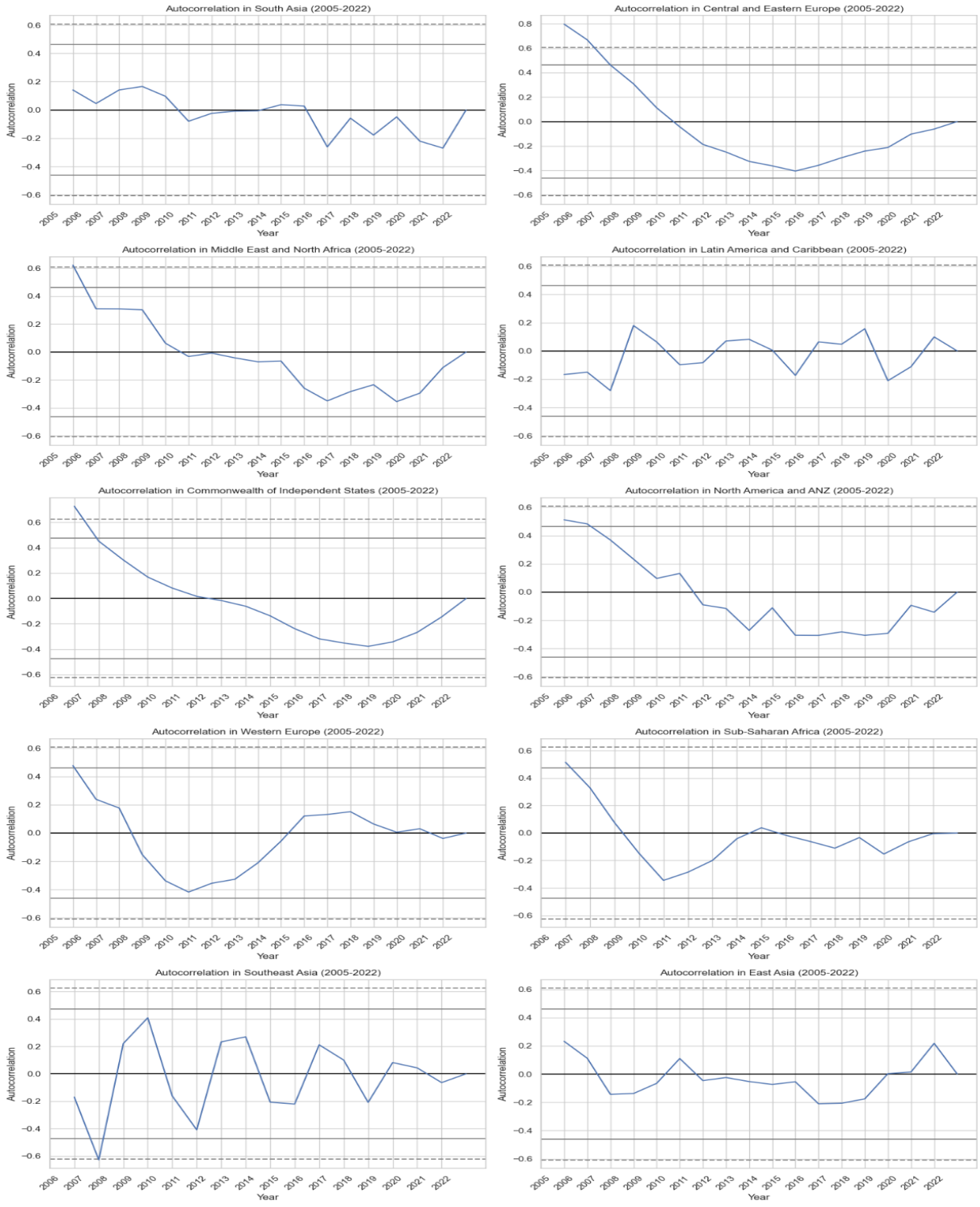


Figure 7: Autocorrelation plots of global regions

Autoregression is applied to the happiness scores of global regions to explore how these scores evolve over time, using 'Year' as the temporal feature. This approach allows the opportunity to uncover potential temporal patterns or dependencies in the happiness scores within each region.

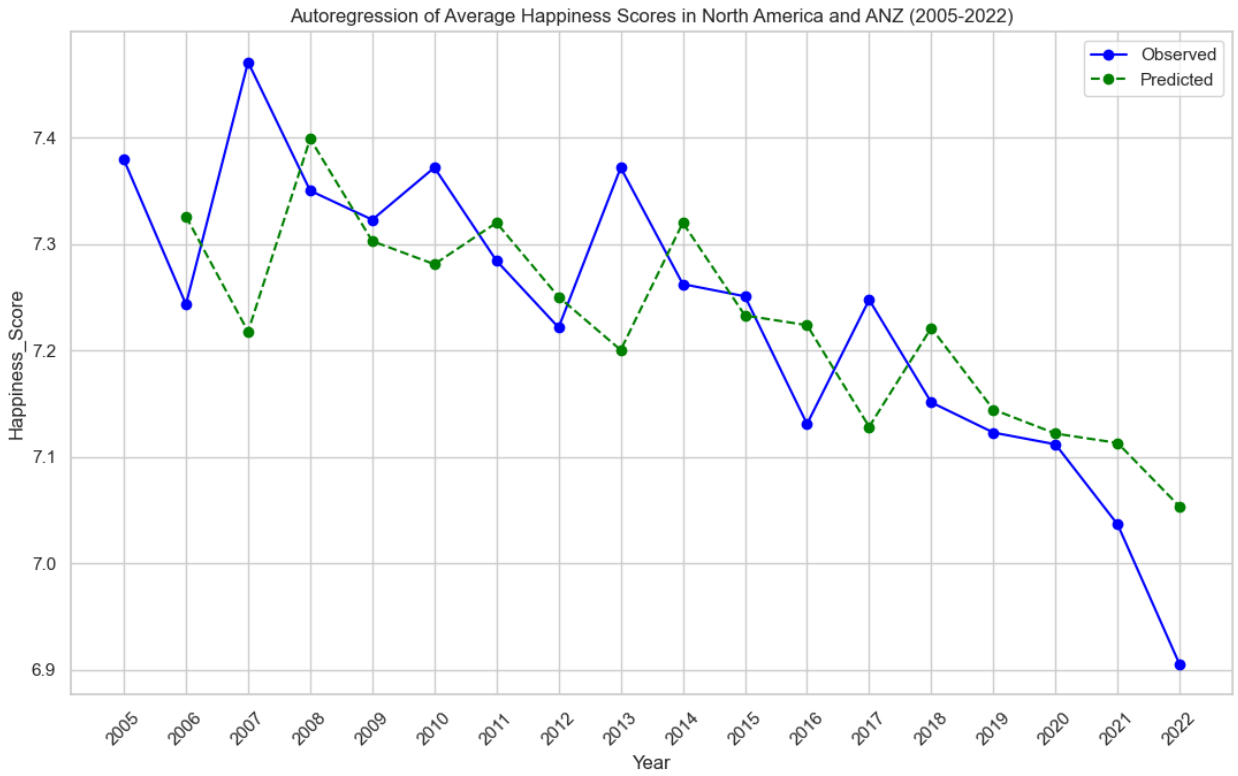


Figure 8: Autoregression plot for North America and Australia and New Zealand (ANZ)

Machine Learning Models

Seven models (Linear regression, Lasso, Ridge, ElasticNet,KNN regressor,Random Forest and XGBoost Regression) were implemented with GridSearchCV for hyperparameter tuning with cross-validation.

Machine Learning Model	Type	Hyperparameters/Values
Linear regression	linear	none
Lasso L1 regularization	linear	alphas:0.001,0.01,0.1,1.10
Ridge L2 regularization	linear	alphas:0.001.0.01.0.1.1.10
ElasticNet	linear	alphas:0.001.0.01.0.1.1.10
KNN regressor	non-linear	n- neighbors: [3,5,7,10] weights:[uniform,distance] p:[1,2]
Random Forest regressor	non-linear	'max_depth': [None, 10, 20], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4], 'max_features': ['auto', 'sqrt', 'log2'] 'n_estimators': [100, 200, 300],
XG Boost	non-linear	Max depth : [3, 10, 30, 100, 300] learning rate: [0.01, 0.1] colsample_by_tree:[0.5,0.7,.09] n_estimators:[300, 500]

Uncertainties due to model splitting can possibly lead to different evaluation metrics. Thus, to account for this cross-validation was used. The evaluation metric used was R-squared because it facilitates straightforward interpretation, representing the percentage of variability in the target variable happiness_score that can be accounted for by the regression models.

4 Results

Since this data set is deterministic there is no standard deviation.
The baseline evaluation for **R-squared is -0.0.**
The machine learning models were chosen for the following reasons:
Linear Regression, Lasso, Ridge, and ElasticNet:

These linear models are chosen for their interpretability and simplicity.
Lasso and Ridge introduce regularization, helping prevent overfitting and handling potential multicollinearity.
ElasticNet combines L1 and L2 regularization, offering a balance between feature selection and maintaining robustness.

K-Nearest Neighbors (KNN) KNN is a non-parametric method chosen for its flexibility and ability to capture complex relationships without assuming a specific functional form. It's particularly useful when happiness patterns of the data do not follow a linear trend.

Random Forest
Random Forest is an ensemble method known for handling non-linearity, interactions, and complex relationships in data.
XGBoost is robust, handles missing values well, and is capable of capturing complex relationships and interactions in the data.

Machine Learning Models Summary

R-squared	**Best alpha**	**Model**
0.80	0.014	Lasso
-1.4	n/a	Linear
0.83	1.0	Ridge
0.85	0.03	ElasticNet

R-squared	**Best Hyperparameter**	**Model**
0.87	0.9	KNN
0.98	Max depth:20,Random min sample:1,n_estimator:150,predicted alpha:0.7	Random Forest
0.91	Max depth:5,subsample:.9,n_estimator:50,learning rate:0.1,	XGBoost

Since the baseline R-squared score was -0.0, any positive R-squared scores obtained from the models indicate an improvement, demonstrating that they contribute positively to explaining the variance in the happiness scores.

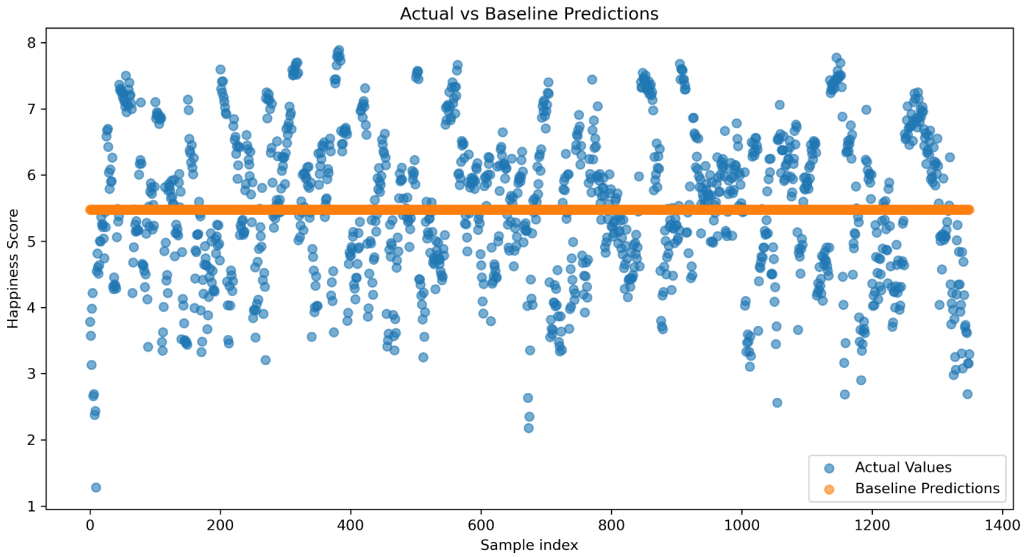


Figure 8 : Actual vs baseline predictions of the models

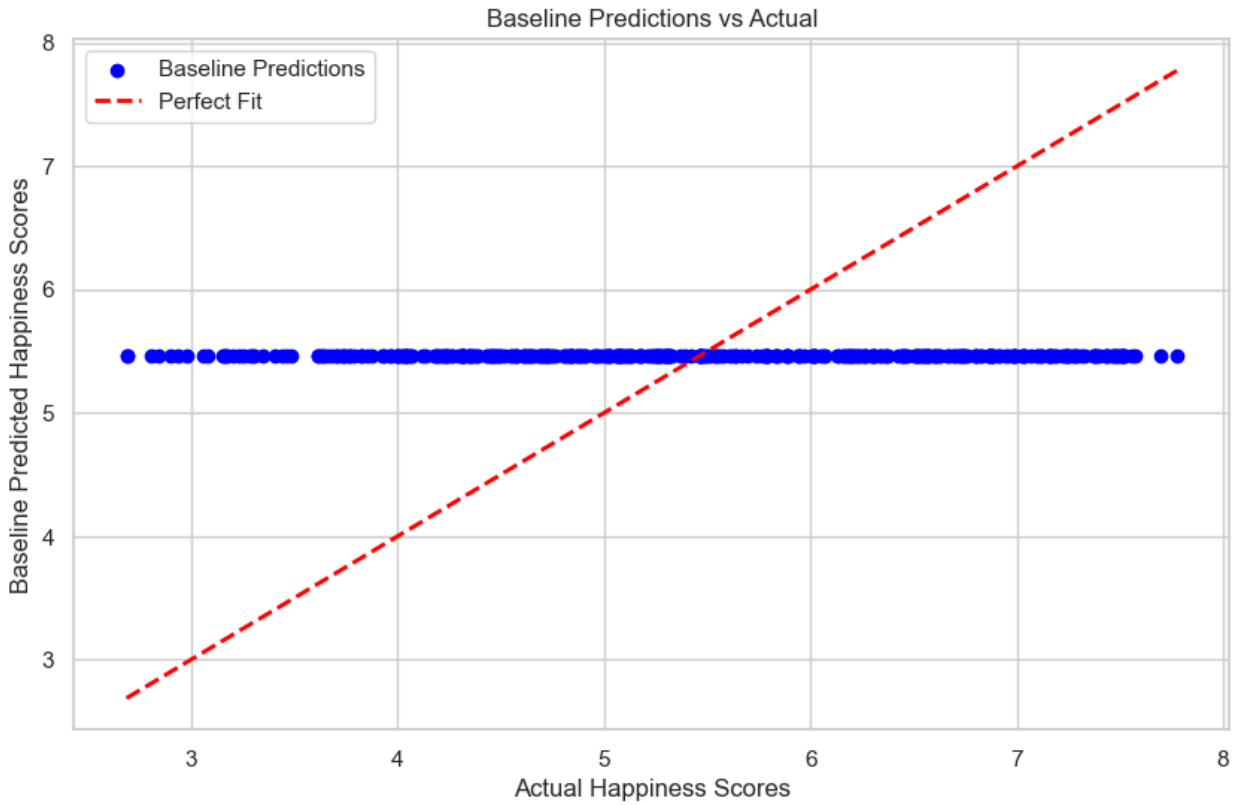


Figure:9 Baseline predictions vs actual

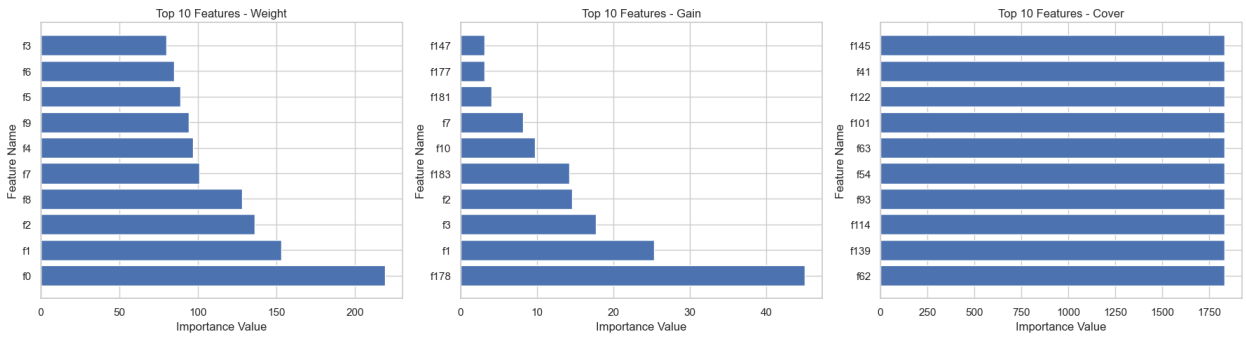


Figure 10: 3 Global features of importance using XGB

Global feature importance using XGBoost in the context of the happiness dataset entails evaluating the impact of each feature on predicting happiness scores worldwide. Features like GDP per capita, social support, life expectancy, and others, contribute to the overall understanding of happiness. The XGBoost model assesses these features, assigning importance based on their influence in capturing the nuanced patterns and factors that contribute to happiness across diverse regions and populations.

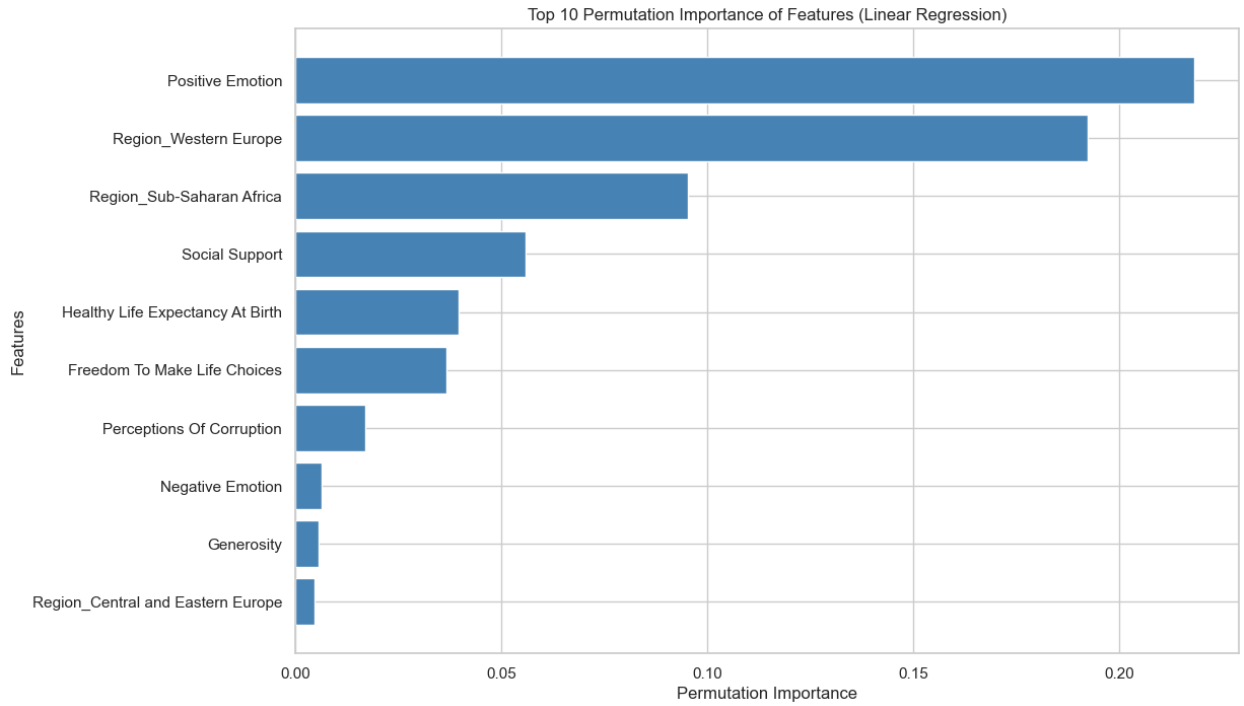


Figure 11: Permutation Importance

In the context of global feature importance, the aspect that most caught my attention was the prominence of positive emotion within the world happiness dataset spanning 2005-2022. This prioritization places positive emotion in a central role, highlighting the importance of factors that contribute to positivity, such as economic indicators, social support, and health metrics.

5 Outlook

To further enhance the performance and interpretability of this machine learning project on the world happiness dataset (2005-2022), employing machine learning models, particularly Linear Regression, offers several avenues for improvement:

1.Feature Engineering explores additional feature engineering techniques to create new meaningful features that might have a stronger correlation with happiness scores.

2. More hyperparameter tuning,which conducts an extensive search for optimal hyperparameters in the Linear Regression model, focusing on parameters like regularization strength (alpha) in the case of Ridge or Lasso regression.

6 References

Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., & Wang, S. (Eds.). (2022). World Happiness Report 2022. New York: Sustainable Development Solutions Network.

Biswas-Diener, R., Kashdan, T. B., & King, L. A. (2009). Two traditions of happiness research, not two distinct types of happiness. *The Journal of Positive Psychology, 4*(3), 208–211.

Gallup (2009). *World Poll Methodology*. Technical Report. Washington, DC.