Introduction and Purpose

Happiness is a crucial measure of societal well-being, reflecting the quality of life and overall satisfaction of individuals within a country. This project aims to explore and model the factors that contribute to a country's happiness, as measured by the Ladder score in the World Happiness Report. The Ladder score, which ranges from 0 to 10, provides a quantifiable metric for understanding the relative happiness of nations worldwide. By leveraging various statistical and machine learning models, this study seeks to identify and quantify the relationship between a country's happiness level and key socio-economic and societal factors such as GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption. The ultimate goal is to determine the most significant predictors of happiness, understand how these factors interact with the ladder score, and build machine learning models and choose a model that best predicts a country's ladder score.

Hypothesis

I hypothesize that factors such as logged GDP per capita, social support, healthy life expectancy, and freedom to make life choices significantly impact a country's happiness score

---

Data Description and Pre-Processing

I used a dataset from Kaggle that includes 149 entries (different countries). The link to the dataset is found here: https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021?resource=download

Data Description

For the purpose of this project, we mainly focused on the numerical features: ladder score, logged GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, perceptions of corruption.

- Ladder Score: A ladder score, also known as a happiness score, is a metric used to measure a country's citizens' happiness.
- Logged GDP per capita: a way to represent GDP per capita using logarithms.
- Social Support: Score on the level of emotional, practical, and informational assistance that individuals within a nation can access from their family, friends, community, and government programs.
- Healthy Life Expectancy: a population health statistic that measures the average number of years a person can expect to live in good health.
- Freedom to make life choices: Score on the right to decide how to live your life and what you want to do with it.
- Generosity: Score on how many people in that country donate money, volunteer, or help strangers
- Perceptions of corruption: Score on how people perceive corruption in their government and other institutions.

Data Pre-processing

The dataset was mostly clean and had no missing values in any rows. In terms of how clean the dataset was, the only error was that the name of the feature for country name was strangely typed so that was changed to simply “Country Name”. With a clean dataset, I would then proceed by changing the objects to numbers using a label encoder. This would then allow me to standardize the data using a standard scaler which would bring all the features to a similar scale. By doing so, we ensure that the features in the dataset contribute equally to the model.

Exploratory Data Analysis (EDA)

Univariate Distribution Analysis

To analyze the distribution of the Ladder Score which is our dependent variable, I created a histogram, a boxplot, and a density plot to visualize the distribution. Across the Visualizations, the Ladder Score displayed a near-normal distribution, centered around the mean with some skewness in the lower tail, indicating that a few countries scored significantly lower on happiness.

Bivariate Analysis

To obtain and visualize the correlation between ladder score and variables of interest, I generated scatter plots for predictors I thought might have a strong correlation with ladder score and then created a correlation matrix and pair plot for the predictor that had the strongest relationship

(Logged GDP per capita) to further visualize it. The visualizations in the bivariate analysis revealed significant relationships between the Ladder Score and key predictors. Logged GDP per capita showed a strong positive correlation ($r = 0.78$), indicating that greater economic prosperity is associated with higher happiness levels. Social support ($r = 0.64$) and healthy life expectancy ($r = 0.66$) demonstrated moderate positive correlations, highlighting the importance of strong social networks and physical well-being in contributing to happiness. Freedom to make life choices also showed a moderate positive correlation ($r = 0.58$), emphasizing the critical role of individual autonomy and freedom in determining happiness. Perceptions of corruption, however, had a weak negative correlation ($r = -0.25$), suggesting that while lower corruption contributes to happiness, its influence is less substantial compared to other predictors.

Multivariate Analysis

A heatmap, pairplot, and PCA (Principal Component Analysis) were implemented to demonstrate the relationships across multiple variables in our data. The multivariate analysis provided deeper insights into how the predictors collectively influence happiness levels, measured by the Ladder Score. Logged GDP per capita emerged as the most significant predictor, showing the strongest positive relationship with happiness, even when accounting for other variables. Social support and healthy life expectancy remained important contributors, with both predictors maintaining moderate positive effects in the presence of others, reaffirming their roles in fostering happiness. Freedom to make life choices also demonstrated a significant positive impact, highlighting the importance of individual autonomy as a consistent determinant of well-being. Perceptions of corruption, though negatively associated with happiness, had a comparatively smaller effect and diminished further when included alongside stronger predictors. Principal Component Analysis (PCA) was used to reduce the dimensionality of the dataset while

preserving as much information (variance) as possible. By transforming the original variables into uncorrelated principal components, PCA simplified the relationships among the factors contributing to happiness, making it easier to interpret the data in a lower-dimensional space. Specifically, PCA identified that the first principal component (PC1) captured 98% of the variance in the data, indicating that most of the information from variables such as Logged GDP per capita, Social support, Healthy life expectancy, and Freedom to make life choices is strongly correlated and can be summarized into a single axis. The second principal component (PC2), which explained only 1% of the variance, added minimal new information. This reduction allowed us to visualize complex relationships and patterns, such as potential clusters or outliers, in a two-dimensional space without losing much detail.

Regression Models

The following regression models were implemented:

- **Linear and Multiple Linear Regression (MLR):** Provided a baseline model to understand linear relationships between predictors and the happiness score.
- **Logistic Regression:** Model to assess classification performance for happiness thresholds. A threshold was added so that there would be a binary result in the ladder score. The threshold was 7 or above indicating a country is happy and below 7 would indicate that country is not happy.
- **Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR):** Models that addressed multicollinearity by reducing dimensionality.

- **Polynomial Regression:** Explored non-linear relationships between predictors and the happiness score to capture complex interactions.
- **Ridge, LASSO, Elastic Net:** Models with regularized regression techniques to improve generalization.
- **Quantile Regression:** Explored the differential impact of predictors across happiness distributions.
- **Cox Regression:** Analyzed survival probabilities for achieving happiness milestones.

K-Fold Cross-Validation

K-fold cross-validation (5 folds) was applied to evaluate model performance more robustly and reduce the risk of overfitting.

Results and Findings

Results across Models

	Model	R^2	RMSE
0	Linear Regression	0.532838	0.694281
1	Multiple Linear Regression	0.621458	0.624969
2	Polynomial Regression	0.516149	0.706573
3	Ridge	0.598459	0.643675
4	LASSO	0.513954	0.708174
5	Elastic Net	0.517983	0.705234
6	PLSR	0.619514	0.626572
7	PCR	0.523566	0.701137

```
Logistic Regression Model Performance Metrics:
Accuracy: 0.83
Precision: 0.29
Recall: 1.00
ROC-AUC: 0.84

Poisson Regression MSE: 0.45
```

QuantReg Regression Results

Dep. Variable:	Ladder score	Pseudo R-squared:	0.5181
Model:	QuantReg	Bandwidth:	0.4866
Method:	Least Squares	Sparsity:	1.357
Date:	Tue, 26 Nov 2024	No. Observations:	149
Time:	20:37:28	Df Residuals:	144
		Df Model:	4

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.8286	0.563	-6.800	0.000	-4.941	-2.716
Logged GDP per capita	0.2439	0.106	2.294	0.023	0.034	0.454
Social support	2.1391	0.816	2.622	0.010	0.526	3.752
Healthy life expectancy	0.0480	0.017	2.900	0.004	0.015	0.081
Freedom to make life choices	2.8437	0.572	4.969	0.000	1.713	3.975

Generalized Linear Model Regression Results

Dep. Variable:	Ladder score	No. Observations:	149
Model:	GLM	Df Residuals:	144
Model Family:	NegativeBinomial	Df Model:	4
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-414.79
Date:	Tue, 26 Nov 2024	Deviance:	1.3522
Time:	20:37:28	Pearson chi2:	1.28
No. Iterations:	4	Pseudo R-squ. (CS):	0.02426
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0885	0.913	0.097	0.923	-1.701	1.878
Logged GDP per capita	0.0494	0.171	0.288	0.773	-0.286	0.385
Social support	0.4498	1.320	0.341	0.733	-2.137	3.036
Healthy life expectancy	0.0062	0.027	0.231	0.818	-0.046	0.058
Freedom to make life choices	0.4738	0.923	0.513	0.608	-1.335	2.283

```
        Current function value: 1.806531
        Iterations: 35
        Function evaluations: 43
        Gradient evaluations: 43
```

ZeroInflatedPoisson Regression Results

Dep. Variable:	Ladder score	No. Observations:	149
Model:	ZeroInflatedPoisson	Df Residuals:	144
Method:	MLE	Df Model:	4
Date:	Tue, 26 Nov 2024	Pseudo R-squ.:	0.04216
Time:	20:37:28	Log-Likelihood:	-269.17
converged:	False	LL-Null:	-281.02
Covariance Type:	nonrobust	LLR p-value:	9.183e-05

	coef	std err	z	P>|z|	[0.025	0.975]
inflate_const	-13.1568	58.935	-0.223	0.823	-128.667	102.353
const	0.0554	0.376	0.147	0.883	-0.681	0.792
Logged GDP per capita	0.0491	0.068	0.721	0.471	-0.084	0.183
Social support	0.4643	0.537	0.864	0.388	-0.589	1.518
Healthy life expectancy	0.0066	0.011	0.608	0.543	-0.015	0.028
Freedom to make life choices	0.4712	0.371	1.272	0.203	-0.255	1.197

Cox Regression Model

model	lifelines.CoxPHFitter
duration col	'time_to_event'
event col	'event_occurred'
baseline estimation	breslow
number of observations	149
number of events observed	17
partial log-likelihood	-39.31
time fit was run	2024-11-27 02:37:28 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Logged GDP per capita	-1.51	0.22	1.11	-3.69	0.66	0.02	1.94	0.00	-1.36	0.17	2.53
Social support	6.00	404.21	11.27	-16.10	28.10	0.00	1.60e+12	0.00	0.53	0.59	0.75
Freedom to make life choices	15.90	8.08e+06	6.93	2.33	29.48	10.25	6.37e+12	0.00	2.30	0.02	5.53

Concordance	0.76
Partial AIC	84.62
log-likelihood ratio test	17.80 on 3 df
-log2(p) of ll-ratio test	11.01

Key Predictors of Happiness

The Quantile Regression model highlighted Logged GDP per capita, Social support, Healthy life expectancy, and Freedom to make life choices as significant predictors of happiness, demonstrating their impact across different levels of the Ladder score distribution. This approach provided a nuanced understanding of how these factors contribute to happiness, particularly in distinguishing their effects on countries with lower versus higher happiness scores. Similarly, the Cox Regression model identified Freedom to make life choices as a critical predictor, significantly reducing the time for countries to achieve higher happiness milestones. These findings emphasize the importance of both socio-economic and individual freedoms in shaping national well-being.

Model Chosen and Justification

Overall, across the various models that would predict the ladder score, which is the score on how happy a country is in 2021, the Multiple Linear Regression (MLR) emerged as the best overall model, with an r-squared of 0.62 and RMSE of 0.62, striking a balance between simplicity, interpretability, and predictive performance. Logistic regression achieved an ROC-AUC of 0.84 and recall of 1.00 but had low precision (0.29), indicating it excels at identifying all happy countries but struggles with false positives. Advanced techniques like Partial Least Squares Regression (PLSR) performed equally well (r-squared = 0.62, RMSE = 0.63) but added robustness to multicollinearity, making it a strong alternative to MLR. Quantile Regression offered a unique perspective by highlighting the impact of predictors across the happiness distribution, revealing significant effects of freedom to make life choices and social support. Specialized models such as Cox Regression identified freedom to make life choices as a critical variable for achieving happiness milestones, though its application was limited to survival analysis. Models designed for count data, including Poisson Regression, Negative Binomial Regression, and Zero-Inflated Poisson, were less effective due to poor fit and limited significance of predictors. Similarly, Polynomial Regression and regularization methods like Ridge, LASSO, and Elastic Net did not outperform MLR. Overall, Multiple Linear Regression remains the most suitable model for predicting happiness scores, given its superior performance and straightforward application. Quantile Regression or Cox Regression complement the findings as they identified the key predictors to determining ladder score as mentioned earlier.

Implementation results of K-Fold Cross-Validation

The multiple linear regression model's performance showed significant improvement after applying k-fold cross-validation. The r-squared score increased from 0.62 to an average of 0.72, indicating that the model now explains 72% of the variance in the data compared to 62% previously. The mean squared error (MSE) also decreased from 0.62 to an average of 0.32, demonstrating a reduction in prediction errors and improved accuracy.

Insights

The findings validate the hypothesis that economic, social, and health-related factors significantly influence happiness. Countries aiming to improve happiness should prioritize policies enhancing these factors. K-fold cross-validation demonstrated the importance of robust evaluation methods to ensure reliable model performance.

Limitations

Despite the improvements in performance from implementing the K-fold cross-validation, the model is subject to several limitations. The multiple linear regression model assumes linear relationships between predictors and the target variable, which may oversimplify complex interactions. Additionally, the presence of outliers or multicollinearity among predictors can still influence the results, even with the use of cross-validation techniques. The dataset itself is limited to countries reporting specific indicators, which may not fully capture global diversity in happiness. Some factors, such as cultural norms or political stability, are not included in the analysis but could play a significant role in determining happiness levels. Expanding the range of

predictors could potentially enhance the model's predictive power. Finally, while the k-fold cross-validation process ensures robust metrics, the findings may not generalize well to different datasets or time periods without retraining and revalidation. As with any model, its applicability to new data requires careful consideration of changes in context and the inclusion of relevant, up-to-date predictors.

Conclusion

Overall, my hypothesis was partially right as some of these factors have a considerable relationship in a country's ladder score. Generosity and perceptions of corruption as seen in their scatter-plots do not seem to have a strong relationship with ladder score. The multiple linear regression model emerged as the most effective model for predicting happiness scores, balancing simplicity, interpretability, and performance. Advanced models like quantile regression and Cox regression provided valuable insights into specific predictor impacts. This analysis underscores the importance of robust evaluation techniques like k-fold cross-validation and highlights the value of socioeconomic and health-related factors in shaping global happiness. For policymakers, this study offers actionable insights to improve well-being through targeted interventions. Future work could expand on this analysis by incorporating additional predictors and testing non-linear models to capture complex relationships.