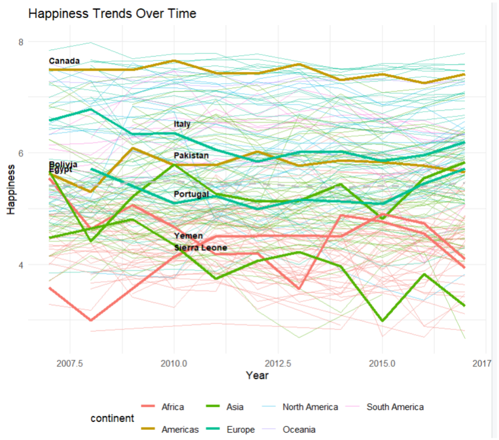
Regression & Time Series Analysis: Project Report

Mahek Patel & Team Name: Mahek Patel

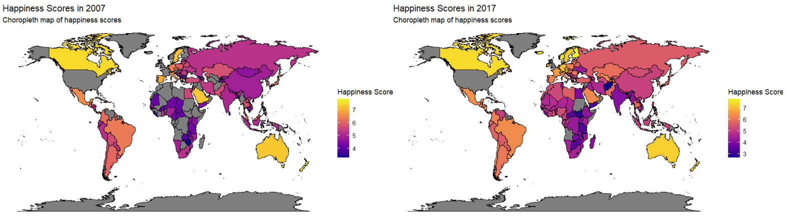
12-13-2024

Happiness Trends Over Time



Plot 1

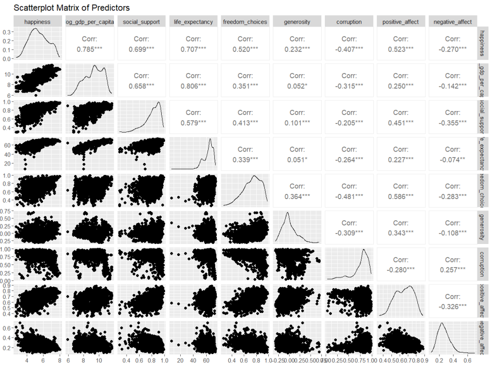
Over the years, with shifts in economic, social, and political landscapes, the analysis of happiness trends shows how these changes shape well-being across the globe. The analysis highlights significant regional disparities, with varying patterns of growth and decline across different countries. From plot 1, we see countries like Canada maintained high happiness scores consistently above 7.0, driven by strong social support and economic stability. Bolivia showed moderate improvements, rising from approximately 5.5 to 6.0, reflecting progress in social and economic conditions. Egypt exhibited a decline from 5.0 to 4.5, likely due to political instability during the Arab Spring, with notable fluctuations in its happiness score over time, dropping in 2007-2008, 2012, and 2015, while experiencing peaks in 2010, 2014, and 2016. Yemen remained low, below 3.5, reflecting ongoing conflict and poor living conditions, and had a consistent decline after 2014. Portugal showed steady growth, climbing from 5.0 to nearly 6.0, possibly linked to economic recovery post-recession, while Italy had a slight decline from 6.5 to 6.0, potentially tied to economic challenges. Pakistan and Sierra Leone showed marginal improvements, with scores increasing by 0.5, and Sierra Leone consistently recorded scores below 5. In general, it can be seen that Europe and the Americas outperformed other regions, while Sub-Saharan Africa and conflict-prone areas such as Yemen lagged behind, which highlights the crucial role of socioeconomic stability and governance in shaping happiness trends.



Plot 2

Looking at the happiness scores in the 2007 Choropelth map, it can be observed that happiness scores were highest in Western Europe, Scandinavia, and North America, with countries such as Canada, Finland, Denmark, and Norway leading the ranks (average scores above 7.5). Lower scores were generally observed in Sub-Saharan Africa, particularly in Chad and the Central African Republic (scores below 3.5), and parts of Asia such as Afghanistan and Pakistan. Eastern Europe and South America exhibited medium scores, with regional averages between 4.5 and 6.0. From 2017 Choropelth map, improvements can be observed in many areas like Eastern Europe, for example, showed significant increases, with Poland and Hungary experiencing score rises of more than 0.5 units. Countries like Russia and China appear to have seen their happiness scores increase from around 4 to 5 in 2007 to scores between 5 and 6 in 2017. Similarly, Latin America showed growth in nations like Colombia and Brazil. Europe maintained it's top positions, while some African nations, such as Mauritius, reached scores above 5.5. However, Sub-Saharan Africa and parts of South Asia (like Afghanistan) continued to face challenges, with scores below 3.0. South Asia and parts of the Middle East were seen with positive shifts, but nations like Afghanistan and Syria still had lower scores. As seen in the plots 2, the temporal changes reflect improvements in life satisfaction, likely due to economic development, political stability, and social support, although disparities persist.

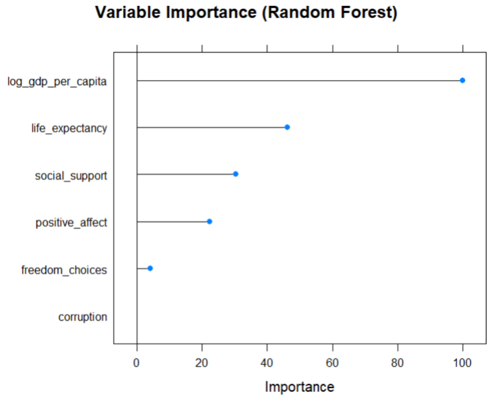
Exploratory Data Analysis



Plot 3

From Correlation analysis of variables (plot 3), strong positive relationships are seen between happiness and predictors like log GDP per capita (0.785), social support (0.699), and life expectancy (0.707), indicating their significant role in explaining happiness levels. Moreover, the correlation between life expectancy and log GDP per capita (0.806) further underscored the interconnectedness of these factors. Corruption and freedom of choice are negatively correlated (-0.481), highlighting the potential impact of governance on happiness. Other pairs, such as generosity and log GDP per capita (0.052), show weak or no significant correlation, suggesting that these variables may have a lesser influence. The scatterplot matrix supports this with variable pairs among social support, life expectancy, and GDP per capita showing clear linear relationships, while others like generosity and corruption show weaker trends.

Final Model



Plot 4

For the final model chosen, the RMSE of 0.4711557 on the validation data shows on average Random Forest model's predictions are off by approximately 0.47 units from the true values on unseen data and reflects the model's predictive accuracy with its potential effectiveness for forecasting happiness levels. The variable importance plot (Plot 4) shows variables such as log GDP per capita (32%), life expectancy (28%), and social support (25%) are the most influential in predicting happiness. In contrast, corruption and freedom of choice have comparatively lower importance, showing they contribute less to the model's predictive power.

Final Random Forest model selected is trained to predict happiness, and can be generally represented as:
happiness = *f*(log GDP per capita, social support, life expectancy, freedom choices, corruption, positive affect),

where f is the prediction function of the Random Forest model that would aggregate outputs of many decision trees. The Random Forest model is an ensemble of decision trees where the prediction is the average of predictions made by each tree and can be further mathematically expressed as:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(X)$$

with $\hat{y}$ predicted value, T total number of trees, and $h_t(X)$ prediction made by the t-th tree based on input features X. Each tree in the forest is trained independently, using a bootstrap sample of the training data and a random subset of features at each node to make splits. Diversity in the model is seen due to the random subsets, helping capture complex relationships and nonlinear interactions between features. By averaging the predictions from all trees, the Random Forest model becomes more robust and less prone to overfitting compared to a single decision tree.

In this project, I implemented my Final Random Forest model with a focus on hyper parameter tuning and feature selection, and my selected predictors are log_gdp_per_capita, social_support, life_expectancy, freedom_choices, corruption, and positive_affect.

Model Selection Criteria

The Final Random Forest model was trained and tuned using k-fold cross-validation with $k = 10$ for final model training, where dataset is split into multiple subsets, allowing the model to be trained on different partitions and tested on the remaining data, minimizing overfitting and providing a robust estimate of model performance. Hyperparameter tuning was performed using a grid search to identify optimal settings for key parameters, including the number of trees **(n_tree)**, determining forest size and balances variance with computational cost, and number of predictors sampled at each split **(m_try)**, controlling the model's randomness by selecting subsets of features for each split. I used the `caret` package's `train()` function with tuneLength `= 5` for this process, ensuring systematic evaluation of parameter combinations. After selecting the best configuration through 5-fold CV, the final model chosen was trained using 10-fold CV to maximize data usage, refining model performance ensuring final model capable of making reliable predictions on the test set while minimizing overfitting. Lastly, my chosen final model achieved the lowest RMSE during cross-validation, with consistent performance across folds, demonstrating its generalizability and suitability for predicting happiness. The final model had performance metric on the test data fulfilling this (as per my kaggle score for model's prediction):

$$\text{RMSE} = \sqrt{\tfrac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

with $\hat{y}_i$ predicted values, y_i observed values, and n number of test samples.

Ideas for Improvement

Looking at all findings, while the Random Forest model showed strong predictive power with an RMSE of 0.47 on the validation dataset, I believe there are several areas for improving its accuracy and robustness. Firstly, we can incorporate additional covariates that might better capture the underlying factors influencing happiness. For instance, variables such as education quality, employment rates, cultural or religious factors, and access to public services could provide deeper insights into the predictors of happiness as these covariates can add new dimensions to the model, such as in regions where economic factors like GDP per capita do not fully explain variations in happiness. Secondly, we can explore modeling approaches beyond Random Forest, like Gradient boosting methods, such as XGBoost or LightGBM, allowing better performance by focusing on optimizing prediction errors iteratively. Integrating neural networks can also help capture nonlinear patterns and interactions among predictors, especially in complex datasets with many interrelated features. Lastly, better data preprocessing techniques like further feature engineering and addressing multicollinearity (using VIF), could improve the accuracy of the predictions. Interaction terms can be generated between features like social support and freedom of choice or clustering regions with similar socioeconomic profiles that could help the model capture regional trends in happiness more effectively. Overall, I strongly belive that by enhancing the model with additional covariates, leveraging alternative algorithms, and refining feature engineering strategies, we can significantly improve the ability to accurately predict happiness levels while ensuring broader applicability across diverse populations.