Machine Learning Bootcamp - 2020

Team No. 4

Project 1 Abstract

Submitted By-
Pratik Ghatake- Business Analyst
Fatmir Likrama- Developer
Celine Cherian – Data Engineer
Miguel Mendoza – Data Scientist

Abstract

Being happy is a simple yet extremely profound feeling nowadays. Happiness cannot be defined in predefined format. How happy are people today? Were people happier in the past? Though it may seem that these questions are difficult to answer or are highly subjective to everyone, we can't deny that these questions undoubtedly matter for each of us personally. In recent years, with the globalized world, it has become even more relevant that governments have started looking at happiness as a matric to measure success. Happiness Index or happiness rank might seem trivial, but it points to the gaping lacunae in the government development policies in each country and one can view the measure as people's perception of how their governments perform. The insights from the analysis of world happiness also helps in pointing out the importance of development qualitatively rather than quantitatively. It takes into consideration economic aspects of the country, socioeconomic aspects such as family contribution to the overall happiness of the person, freedom, health etc.

This project focuses on predicting the happiness score of different countries over time. Also, another goal would be what major factors have caused the countries rank to increase or decrease. Additionally, findings like did any country experience significant change in the happiness score. We will also investigate what are most informative attributes and how much they contribute to the prediction model. This data will allow us to analyse the happiness score of different countries for different years and will enable us to investigate the ways that make people well-being. Correlation among the features will also be carried out in order to verify if the features themselves are not highly correlated. We will narrow down the key variables which contribute the overall happiness score.

The outcomes and the insights in the data can help us decide which countries are best to do business with. The analysis can be used by researchers to dig down in detail what factors of the government can help to increase the happiness score of the country? What change or policies if improved or changed can help in gaining more score in the health category. In the highly dynamic and globalized world, the analysis can be used as an input to business plans to add value.

Data Preparation and Analysis

Dataset Resource-

World Happiness Report Datasets: https://www.kaggle.com/unsdsn/world-happiness

Data Pre-processing for Machine Learning: https://www.kaggle.com/getting-started/167283

Data Cleaning-

- To make the dataset compatible for the analysis, we first wanted to check if the data has missing values. We went through 5 datasets and determined that there are no missing values.
- Next, we dug deeper into the data and found that data is not consistent. The features in different datasets have different variations.
- In order to proceed ahead for data integration phase, we renamed some of the features and maintained the uniformity in the feature's names. The final features after renaming are as follows,
    1. Country or Region
    2. Happiness Rank
    3. Happiness Score
    4. Economy (GDP per Capita)
    5. Social Support
    6. Health (Life Expectancy)
    7. Freedom to make life choices
    8. Generosity
    9. Perceptions of Corruption
- We will also be checking for the multicollinearity in the data by using Variance Inflation Factor (VIF) test. The multicollinearity in the explanatory variables must be avoided in order to have better accuracy of algorithm since it undermined the statistical significance of independent variable.
- The dimensionality reduction possibilities will be checked. In other words, we will be checking if we can potentially integrate some of the variables to form more informative features.

- The target variable for our analysis will be "**Happiness Score**".
- For descriptive analytics, we will be using Matplolib, Seaborn libraries to present the summary of facts in understandable format for further analysis.

Hypothesis

The Multiple linear regression will be applied on the data for predicting "Happiness Score". Ridge and Lasso Regression will be used to check for prediction error as well as multicollinearity. The outputs from Ridge and lasso regression will be compared to select the one that produces the best fit using sklearn's mean absolute error function to compare our prediction data with the validation data. We will be taking input as an instance of machine learning model and then most informative attributes will be decided by Recursive feature elimination method.