**ARE YOU HAPPY?**

BY – ADITYA BHATT

**ABSTRACT**

We live in times where stress rules our mind. Often it gets ignored and no one pay’s attention

To. We try to model ladder score (that’s indicating how happy you are) using regression based approach. The final model that was build use two features one economic **(LOG GDP PER CAPITA)** and other non economic **(PERCEPTION OF CORRUPTION) .** Random forest regressor was used and we got a accuracy of about 70%.

In the due course of the paper the reader will get a understanding of how and why modeling Happiness index ( in the form of ladder score ) .Gives country an idea of how they can improve their ladder score to top the global ranking and make their citizens happy.

The Link to the final app –

<https://happynessindex.herokuapp.com/>

**INTRODUCTION**

The complete life cycle of a data driven project was followed.

**Data Collection**

Data is always a key component to any analysis. The data set was link from the following link –

<https://worldhappiness.report/ed/2021/#appendices-and-data>

The data was taken from a official report which ranked countries on the basis of happiness index

Finland was the most happy country.

**Data Dictionary (key features)**

In the official report lot of features and many sources for different years were fetched and import features were chose which were –

Log GDP per capita, Social support, Healthy life expectancy at birth, Freedom to make life choices, Generosity, Perception of Corruption

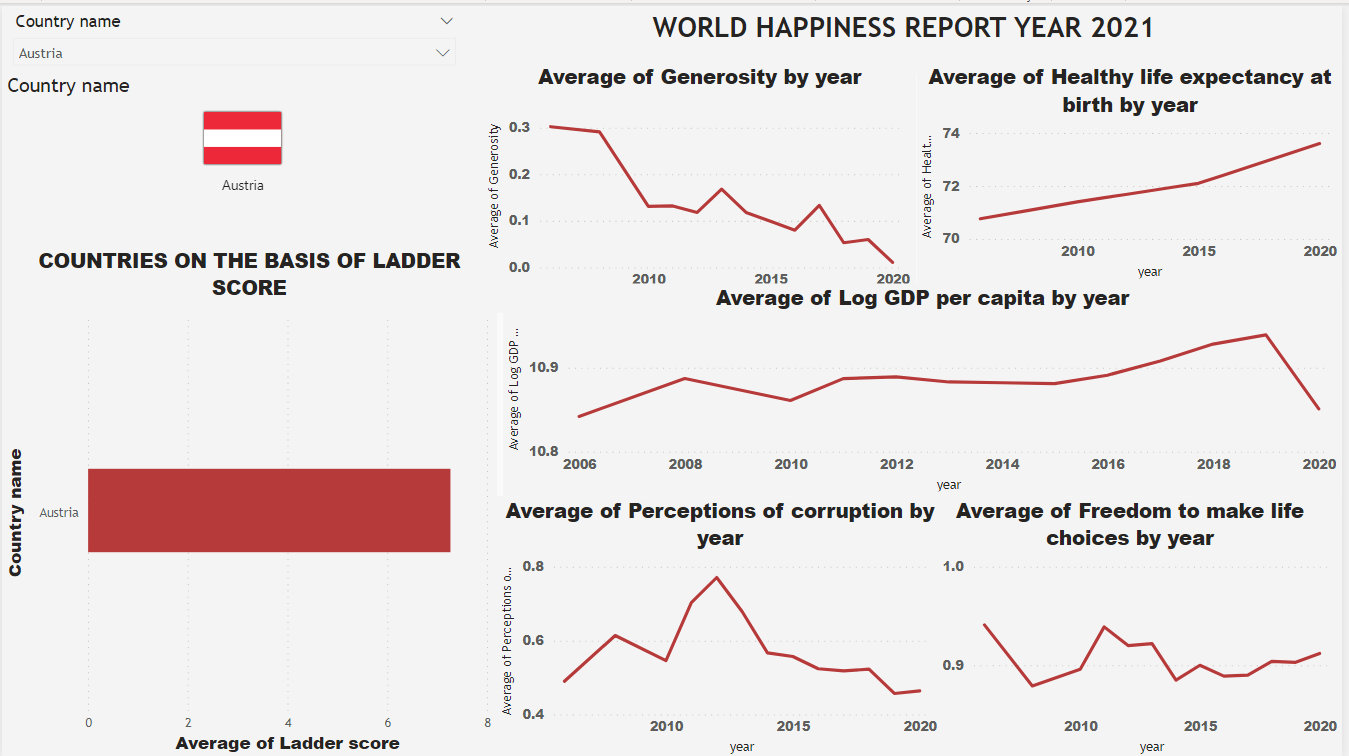
These features are calculated by data collection team at world economic forum so we have relied on their data and way to compute / collect data.

**Data Cleaning**

The first step was to drop the null values since they were less than 10% of the total data and since the data is computed in a different way so computational techniques wouldn’t have been a logical approach.

**DASHBOARDING / VISUALIZATION –**

A complete dashboard was created with had all features country wise. One could use this dashboard to present and analyze country wise insights and help a country view where they lack and hence therefore they can work on the sector and improve their global ranking.



**Figure (1)**

This helps us to view the analysis country wise and thereafter analyze the results to infer where exactly the country lacks.

**EDA AND FEATURE SELECTION**

When we target a regression project eda and feature selection is an integral part of the project since a lot of assumptions are involved.

So were weren’t sure that whether a linear or a non linear model will be used. But later on we used a random forest regressor which is something beyond and has a different working altogether.

So starting with **Normality** some features such as Social support , Health Life Expectancy and freedom to make life choices showed some deviations from the normal behavior. The skewness was expected since they are economic variables a lot of transformations log x ,1+log x, square root transformations were used but much changes weren’t seen.

Then **yeo johnson transformation** was used and we got the desired results but transformed might have good results but it’s interpretation remains a concern. But we kept the original dataset as well as the actual dataset with us for future benchmarking.

Then we jumped into **Multicollinearity** so before starting with multicollinearity we always knew that the features here are highly correlated.This kind of problem is common in economics data.

Now we plotted a correlation matrix and heatmap that helped us confirm our assumption that the data suffers from correlation.

Then **VIF** based analysis was carried out this helped us analyze how VIF came out to very much greater than 5 for almost all variables. But when we carried out VIF analysis on the transformed data the results were amazing almost all features had VIF < =5 but since starting we were skeptical about the transformed data. So we got know that our data suffers from multicollinearity and transformed data is performing way better than normal data in context to multicollinearity

As a remedy based solution we used **PCA** and using PCA (features =2) on the real dataset obtained a accuracy of 96 % (using random forest regressor ) but we were didn’t want to use PCA since we had a goal of creating app and during testing a single record is passed and extracting features from a single record doesn’t make sense . But PCA testing Batch testing may be a solution but we didn’t wanted to

go that far.

So until know the problem multicollinearity wasn’t solved.

We went to **Heteroscedasticity** we first conducted residual analysis using graphical method we could see that there is no severe heteroscedasticitybut this illusion was proven wrong when we conducted

**Breush Pagan test and NCV** test which proved that we have heteroscedasticity both in actual dataset and transformed dataset.

We decided to drop features we decided to go with one economic variable and other non economic variable i.e **LOG GDP PER CAPITA** and **PERCEPTION OF CORRUPTION.** We went by this approach since we believed that features are highly correlated and indirect correlation is also seen.

Now there was no multicollinearity in the data, no heteroscedasticity in the actual data.

No multicollinearity was there in the transformed data but heteroscedasticity was there.

So the our fear got proven right and we concluded that **“REMIDY IS DANGEROUS THAN THE DISEASE ONLY”** so at last we decided to go with the above mentioned two features.

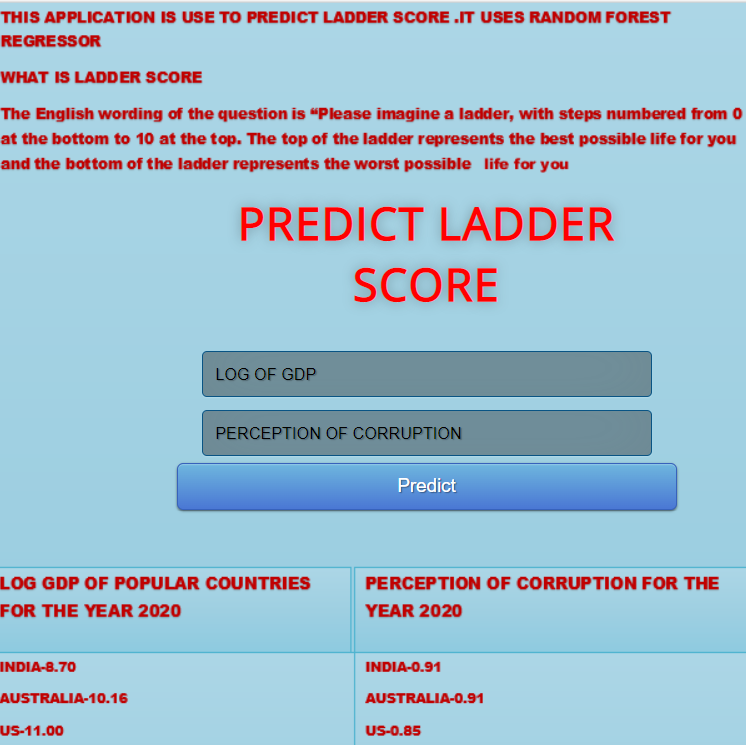
**MODEL TRAINING**

The model training was done using random forest regressor and we got a accuracy of 70%.

The no of estimators chosen after hyperparameter tuning was 100.

**DEPLOYMENT**

A front end was created using html and css. Linked with regression app using FLASK API. And the application was deployed on Heroku (cloud platform).



**FIGURE -2**

**This app can be tested and used by all users.**

**REFERENCES –**

**1.** <https://worldhappiness.report/ed/2021/#appendices-and-data>

**2.** [**https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/**](https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/)

**3.** [**https://www.r-bloggers.com/2016/01/how-to-detect-heteroscedasticity-and-rectify-it/**](https://www.r-bloggers.com/2016/01/how-to-detect-heteroscedasticity-and-rectify-it/)