Data science seminar – Final submission

X

Yosi Amiga

Itay karat

Steps in our project:

1. Data pre processing
2. EDA – explanatory data analysis
3. Machine learning models for evaluating correlation and analysis of feature importance
4. Business outline

World happiness report data analysis and collection goal over-view:

The collection of World Happiness Report data serves several important purposes:

- Comparative Analysis: By collecting happiness data from various countries over time, the World Happiness Report allows for a comprehensive comparative analysis of well-being and happiness levels across different nations. This helps policymakers and researchers identify patterns, trends, and factors that contribute to happiness or its absence in specific regions.

- Policy Evaluation: The data collected in the World Happiness Report provides a valuable tool for evaluating the effectiveness of various policy interventions aimed at improving well-being. Governments and organizations can assess the impact of policies related to education, healthcare, employment, social welfare, and other domains on the happiness of their populations, enabling evidence-based policy decisions and adjustments.

- Public Awareness: The report helps raise public awareness about the importance of happiness and well-being as key indicators of societal progress. By disseminating the findings widely, it encourages individuals, communities, and governments to prioritize holistic well-being and pursue policies that promote happiness.

- Targeted Interventions: The data from the report can guide targeted interventions in specific regions or communities that have lower happiness levels. It enables policymakers to identify the underlying factors contributing to low well-being and design interventions that address those specific challenges, such as improving access to healthcare, reducing income inequality, or enhancing social support systems.

- Longitudinal Monitoring: The collection of data over time allows for longitudinal monitoring of happiness trends. This helps identify whether happiness levels are improving or declining in different countries and regions. It provides an opportunity to detect emerging issues or challenges to happiness and well-being, facilitating proactive measures to address them promptly.

In summary, the collection of World Happiness Report data serves as a valuable resource for comparative analysis, policy evaluation, public awareness, targeted interventions, and longitudinal monitoring, ultimately contributing to global efforts to enhance happiness and well-being for all.

Data overview:

- **Some of the features included in the report (years: 2015-2019)**

Country: The name of the country for which happiness data is recorded.

Year: The year in which the happiness report was published or the data was collected.

Happiness Score: A numerical measure representing the overall happiness level of a country, often based on various factors and surveys.

GDP per capita: The Gross Domestic Product (GDP) per person, which is a common economic indicator used to assess the wealth and standard of living in a country.

Social support: The extent to which individuals in a country have social connections, social networks, and support systems.

Life expectancy: The average number of years a person is expected to live in good health, reflecting the quality of healthcare and overall well-being.

Freedom to make life choices: The degree of individual freedom and autonomy in decision-making, personal rights, and civil liberties.

Generosity: The tendency of individuals in a country to engage in charitable acts, help others, or donate resources.

Corruption Perception: A measure of the perceived levels of corruption in government and public institutions.

Dystopia Residual: A hypothetical benchmark representing the lowest possible happiness score, often used to compare and contextualize the happiness levels of countries.

Number of Countries: The number of countries included in the dataset varies each year, with approximately 150 countries covered in each year's data.

Economics data set Features (2015-2019):

- Access to clean fuels and technologies for cooking (% of population)
- Access to electricity (% of population)
- Agriculture, forestry, and fishing, value added (% of GDP)
- Imports of goods and services (% of GDP)
- Industry (including construction), value added (% of GDP)
- Manufacturing, value added (% of GDP)
- Military expenditure (% of GDP)
- Tax revenue (% of GDP)
- Number of Countries: Data available for most countries globally, but only the ones present in the Happiness Report were considered for the final merged dataset

Data state analysis:

Handling missing values:

- We used a simple solution using the simpleImputer API in python (for forward work we can try other imputer strategies and figure out if there's any difference):

```
numeric_columns = merged_data.select_dtypes(include=['number']).columns
mean_imputer = SimpleImputer(strategy='mean')
merged_data[numeric_columns] = mean_imputer.fit_transform(merged_data[numeric_columns])
```

Forward to be experimented are some other imputing methods such as:
- regression imputation
- k nearest neighbors
- Random sample imputation
- Deep learning based imputation

World happiness report:

- Missing Values: Some features, such as 'Trust (Government Corruption)' and 'Generosity,' have missing values for certain years and countries. These missing values were addressed during the data pre-processing stage.

Economics data set:

- Missing Values: Some missing values were observed, which were handled during data pre-processing.

Cleaning and Filtering:

- Some of the features in the tables has the same effect on the model performance, as we studied that might be because of collinearity or some other domain knowledge reason.

  In our case we decided to drop some of the features in order to maintain a lean accurate model.

  For example, 'Happiness Rank' was dropped for being in a high correlation with the target 'Happiness Score' we wish to predict.

  We only included countries that appeared in all datasets to ensure complete and accurate analysis.

- Another method for transforming the data as part of the pre-processing stage is: MinMaxScalar.

  - 'Year' - from 2015 to 2019, 'Happiness Score' - from 0 to 10 and more.

- We also **dropped columns and rows with missing values exceeding a certain threshold** (70% in this case).

```python
# Set a threshold for the percentage of missing values you are willing to accept
threshold = 0.7

# Drop columns with more missing values than the threshold
missing_column_ratio = merged_data.isnull().mean()
merged_data = merged_data.loc[:, missing_column_ratio <= threshold]

# Drop rows with more missing values than the threshold
missing_row_ratio = merged_data.isnull().mean(axis=1)
merged_data = merged_data.loc[missing_row_ratio <= threshold, :]
```

- Forward work is to add outliers removal.
  - In order to do so it is more likely recommended to speak to some domain expert that will explain the reasonable distribution of the features values.

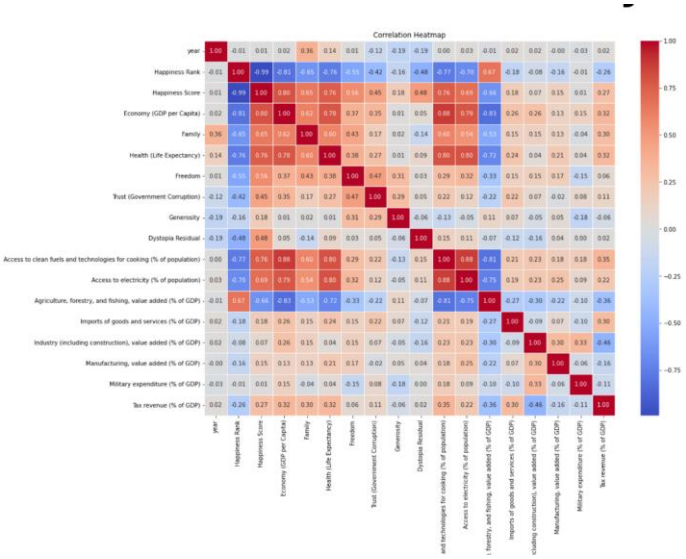
  what is a normal value and what is an outlier of some feature.

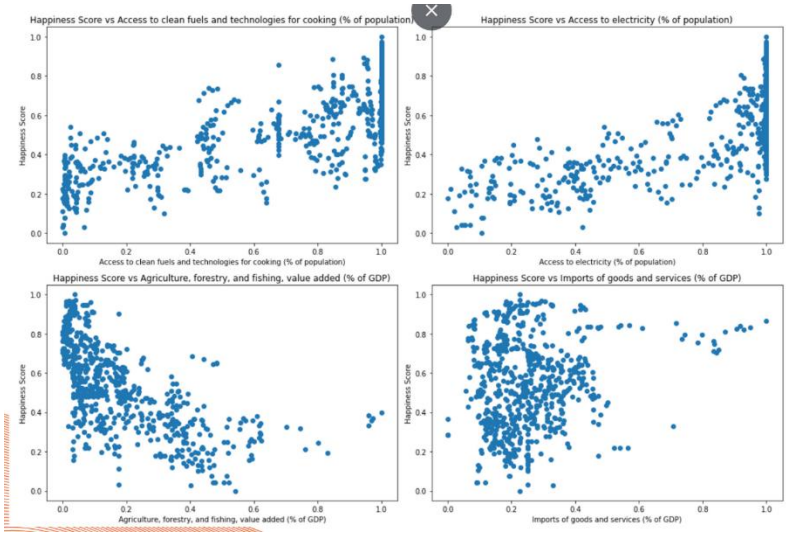
  Exploratory Data Analysis:

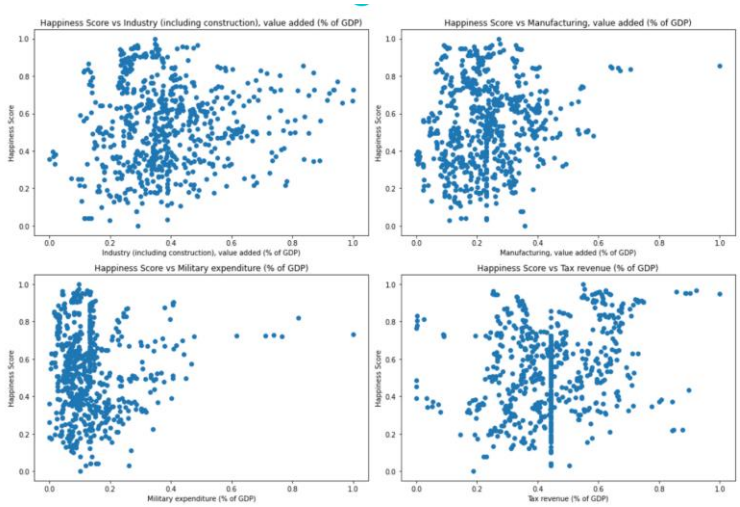
- In order to get a better feel for the data we performed some EDA process and it helped us to learn:

  - Correlations
  - Patterns
  - Density of some features and it's distribution

  We tried to use as many visualizations as we can to continue to the next steps in the pre processing and feature selection process.







Final step:

* Experimenting different ML regression models.

- In this part we conducted a series of experiments in order to create the most accurate regressor we can.
- Different configurations were conducted to fine tune the model and get better results.

Models we tried:

- Random forest
- Decision tree
- Linear regression
- XGboost

Forward to be examined:

- Other boosting methods
- Different deep learning methods

- **Pipeline:**

  Train a **regressor** on the original dataset without:

  1. Happiness Rank

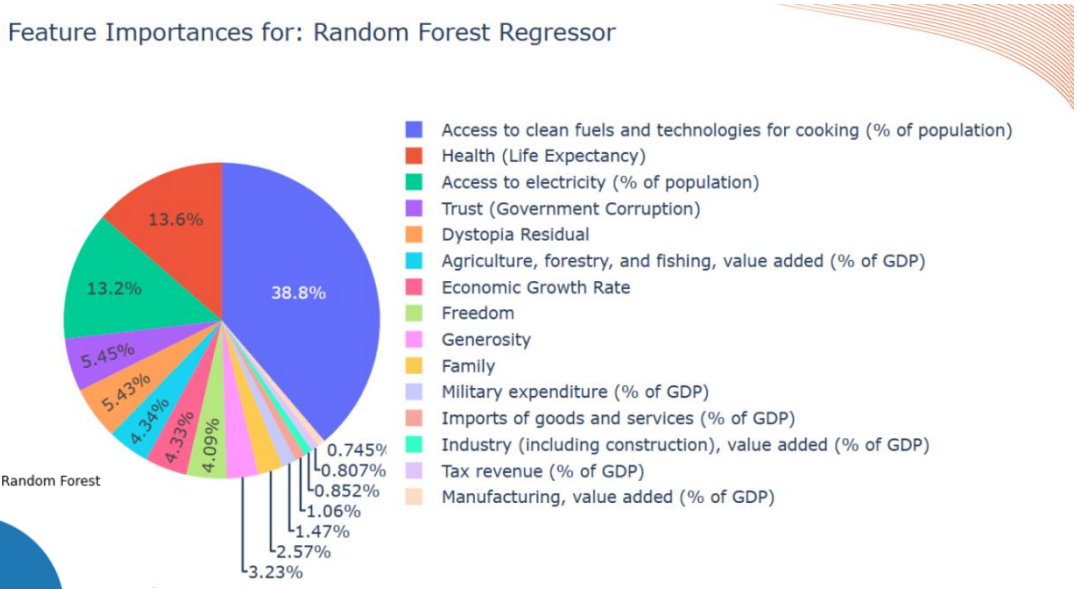
  2. Economy (GDP per Capita)

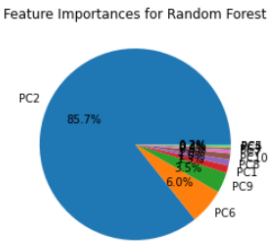
  3. Happiness Score

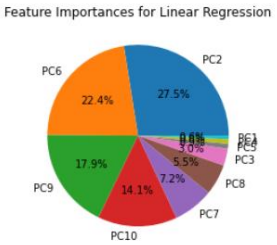
  4. Countries

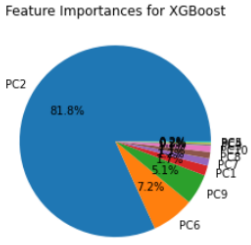
  5. Years

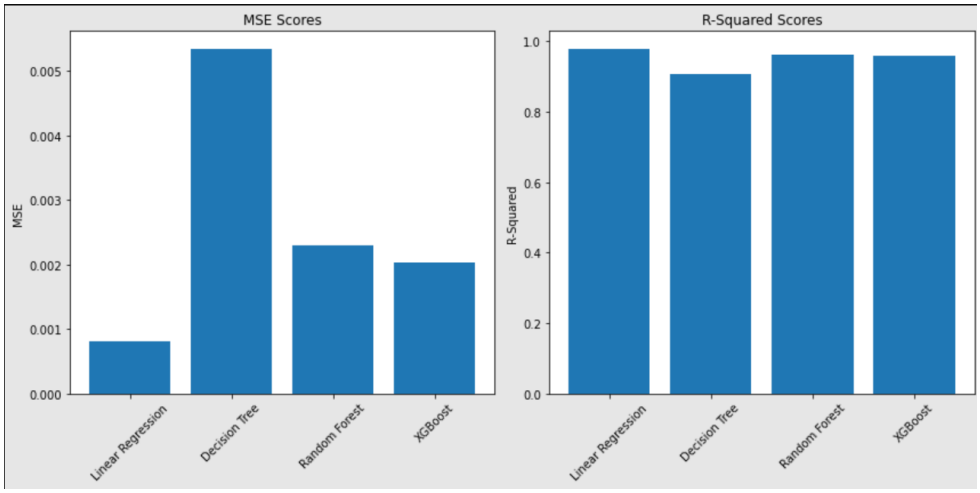
For each model we can create the Feature importance plot and try to explain the model output and maybe find intersection between important features found be several regressors.

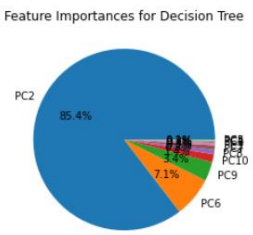












Conclusions:

The goal of our project was to model and understand the key drivers of happiness levels across various countries using different machine learning methods: Linear Regression, Decision Trees, Random Forest, and XGBoost.

When interpreting the coefficients for the Linear Regression model, it's important to remember that they represent the change in the dependent variable (happiness level) for each one unit change in the predictor, assuming all other variables are held constant. For example, for each one unit increase in 'Dystopia Residual', we can expect an increase in happiness level of approximately 0.62 units, while keeping all other predictors constant.

Comparatively, the Decision Tree, Random Forest, and XGBoost models indicate feature importance rather than direct coefficients. These indicate which variables are most influential in predicting happiness level, based on their usage in creating splits in the decision trees.

In this project, it appears that "Access to clean fuels and technologies for cooking (% of population)" was a common significant predictor across all models.

Variables like "Health (Life Expectancy)", "Dystopia Residual", and "Generosity" also appeared frequently as top predictors. It is worth noting that economic variables like "Economic Growth Rate", "Military expenditure (% of GDP)", and "Tax revenue (% of GDP)" were generally found to have lower influence in these models, contrary to what one might intuitively expect.

The performance of our models was evaluated using Mean Squared Error (MSE) and R-squared statistics. A lower MSE indicates a better fit of the model to the data, while a higher R-squared value indicates that the model explains a larger proportion of the variance in the dependent variable.

In conclusion, this project provided insightful findings on the key predictors of happiness levels across different countries. Future studies may consider exploring other machine learning algorithms, adding more variables to the models, or using different strategies for handling missing or categorical data to further improve the performance of the models.