Intro to Data Mining taught by Amir Jafari, PhD

The economics of happiness

Proposal

Lilian Sao de Rivera
4-6-2019

Table of Contents

PROPOSAL: The Economics of Happiness

Introduction

Retiring in the United States is getting harder because of the costs of living. Sites that help people to plan for the future usually states that it is necessary to have one million dollars to live comfortably when people retired. Many Americans do not have enough money to cope with an emergency even if they are educated. The debts that Americans carry may be high due to education loans, house loans, and health expenses. However, some Americans have coped with these problems at the end of their lives by looking for cheaper places to live, where their savings and retirement income can support them a good life. However, these locations may be outside of the country. Before taking any decision, the average American that is willing to take this decision should have information about these locations. This type of decisions not only need to be related to retirement but also about buying a vacation home or having a gap year in an exotic place.

This type of decisions has to do with wellness, and for this reason, the index of happiness is a good indicator to choose a country. If someone decides to go on an adventure, wellness, security, and economic stability are good indicators. The index of happiness created by the Canadian Institute for Advanced Research measures the level of happiness using a poll that measures several aspects of social life. The GDP is the only economic indicator, and the perception of government trust is used, but it should be highlighted the word perception. Thus, there should be another way to predict this happiness. GDP should be kept on the analysis, but other features are included like, the economic group of each country, the GINI indicator, that is the gap between rich and poor people. Is a country with a big gap happy? Do worldwide governance indicators contribute to well-being and happiness of the country? Indeed, good indicators related to economics will allow anyone to analyze the economic stability of a country but also happiness. If the country is happy and then the related features about economic stability are right, then a good investment can be made.

Therefore, it can be stated that this project looks into the relationship between these indicators and happiness as a primary goal and as a second goal looks into the contribution of these variables into the final index. If a country has a high level of happiness determined by economic factors; thus, it is a viable place where a person can retire or buy a vacation home that would be a good investment.

Data

The information to classify the level of happiness is contained in four datasets.

Happiness index found in Kaggle web site.

This dataset contains:

- Information of the year 2017
- Happiness index from 156 countries

The dataset contains several columns, like GDP and family, that state a ratio of contribution. These additional columns will not be used in the final algorithm. The happiness index is present in all the observations. The following datasets will be subset based on the number of countries that the happiness index presents.

Governance indicators from the World Bank and the Brookings Institution.
This dataset contains:

- Information from the years 1996 thru 2018
- Indexes from Voice accountability, Political Stability, and Absence of Violence/Terrorism, Government Effectiveness, Regulatory Quality, Rule of Law, and Control of corruption.
- Information from 216 countries

This dataset is an excel with a worksheet for every indicator. The final dataset needs to gather all the sheets in one dataset with the right index. The index by country has missing values. The approach will be to copy the last index into the blank spaces. Some indexes are taken every four or five years, so it will be assumed that all the conditions stay the same and the index only will be translated from year after year until new data is found. Finally, only the countries that are present in the happiness datasets will be filtered for the year 2017.

GDP indicators from the World Bank
This dataset contains:

- Information from the years 2006 thru 2017
- GDP indicator for 264 countries

This dataset has information for almost all 264 countries. Nevertheless, there some countries that do not have information for 2017. For these countries, the GDP from the last year with data will be copied into the missing cells. Then the countries will be chosen based on the first dataset.

GINI indicators from the World Bank
The dataset contains:

- Information from the years 2006 thru 2017
- GDP indicator for 264 countries

This dataset has information for almost all 264 countries. Nevertheless, there some countries that do not have information for 2017. For these countries, the GDP from the last year with data will be copied into the missing cells. Then the countries will be chosen based on the first dataset.

Finally, all four datasets will be joined by country. Only the year 2017 will be taken from each dataset.

Methodology

Once the data is cleaned, filtered and joined, the index of happiness will be binned into four groups. The groups will be 'Happy countries,' 'Medium happy countries,' 'Low happy Countries,' 'Not so happy countries' and the purpose of the data mining algorithm will be to predict the group where it belongs base on the governance, GDP, and GINI indexes. The chosen algorithm will be a Decision Tree, using Random Forest to get the best results. The other purpose of the research is to look for the features that contribute more to the happy index. The random forest from Scikit-learn provides methods to obtain this information. To this point, it seems that the standard algorithm from Scikit-learn is enough for the present research if some additional changes should be made they will be addressed in the final report.

The software that will be used is phyton. The libraries that will be expected to used are pandas and Numpy to clean the information, and for EDA, Matplot and Seaborn for EDA, Scikit-learn for the decision trees, random forest and metrics for the data mining, and finally PQTy will be used to tie everything together in a data product for the public consumption.

The dataset about the index of happiness has a companion document "Word Happiness Report 2017" which address important questions about what information was used to calculate the indexes. This research will only sue GDP within the same features, other indicators will be included since the main purpose of the project is for investment purposes there is necessary to include economic factors that make sense for an investor. The Brookings Institution has an interesting paper about inequality and the causes of inequality on hope. A good GDP not necessary mean that a country is in good shape since it is an average, which means that an increase of poverty can be obscured in the final results by the increment of wealth in corrupt countries for example. Thus, the GINI factor and indexes of corruption are included along with these studies to justify their presence in the study. For the technical part of the paper the book "Python Data Science Handbook" by Vanderplas, J., and notes from the class Intro to Datamining by Jafari, A. will be used.

Evaluation

To evaluate the effectiveness of the model the Scikit-learn metrics library will be used, specifically the MSE of the model with the Random Forest method. The same way ROC and AUC will be used to determine how well the model predict base on the features. There will be using the confusion matrix to see visually in numbers and graphics how well the model performs, so the final product and result can be easy to understand.

Schedule

The proposed schedule for the project is as follows:

Final Project : The Economics of Happiness		
Phase	**Activity**	**Due Date**
1. Definition of Problem	Defining the problem	4/6/2019
	Documentation	4/6/2019
	Data sets	4/6/2019
	Definition of goals	4/6/2019
	Presentation of proposal	4/6/2019
2. Preprocessing	Cleaning Datasets	4/8/2019
	Transforming Information	4/8/2019
	Final Data Sets for model	4/8/2019
3. Data Mining and Evaluation	EDA (Statistics and Graphics	4/13/2019
	Creation of Decision Trees	4/13/2019
	Creation of Random Forest	4/13/2019
	RMES	4/13/2019
	Confusion Matrix	4/13/2019
	AUC	4/13/2019
	Contribution of Features	4/13/2019
4. Creating Deliverables	PQTY for the data product	4/20/2019
	Creation of Final Report	4/23/2019
	Introduction	4/23/2019
	Methodology	4/23/2019
	Presentation of Results	4/23/2019
	Conclusions	4/23/2019
	Recommendations	4/23/2019
	Creation of presentation	4/23/2019
	Creation of Individual report	4/23/2019
	Creation of Github	

References

Graham, C. (2017, December 1). *The human costs of the productivity paradox in the USA: Insights from metrix of well-being.* Retrieved from Brookings: https://www.brookings.edu/research/the-human-costs-of-the-productivity-paradox-in-the-usa-insights-from-metrics-of-well-being/

Helliwell, J.F., Layard, R., Sachs, J.D. (2017). Word Happiness Report 2017. *Canadian Institute for Advanced Research.* Retrieved from https://s3.amazonaws.com/happiness-report/2017/HR17.pdf

Kaggle. (2017). Global indicators happiness. Retrieved from https://www.kaggle.com/dgscharan/data-set-for-happines

VenederPlas, J. (Dec, 2016). Python Data Science Handbook. *O'Reilly Media Inc*.

World Bank.(2017). GDP per country. *World Bank Data.* Retrieved from

World Bank.(2017). GINI per country. *World Bank Data.* Retrieved from

World Bank.(2018). World Wide Governance Indicators. *World Bank Data and Brookings Institution.* Retrieved from http://info.worldbank.org/governance/wgi/index.aspx#reports