ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

University of Zurich UZH

Prediction: Happiness in Kazakhstan

Mirjam Grünholz
Master in Comparative and International Studies
Department of Humanities, Social and Political Sciences, ETH Zürich

1 Introduction

National success is often measured by GDP and is set equal to well-being. This is misleading because GDP does cover monetary values and market transactions but ignores values such as health, education, cultural or ecological diversity. Happiness is a widely discussed alternative measure to GDP which tries to capture well-being and life satisfaction. Therefore, this poster aims at fining a good statistical model to predict happiness. The various models are tested on World Value Survey data from Kazakhstan, a multi-ethnic country with rich traditions. 88% of the respondents reported to be happy or rather happy, which is 4% more than the world average.

2 Data Preparation

Data Used

- Data: World Value Survey, wave 6, 2010-2014
- Country: Kazakhstan
- Dimension: 1500 rows, 440 columns

Data Reduction

- All negative values (i.e. not asked, don't know) are changed to NA.
- All variables with more than 20 percent NAs are deleted.
- Variables with values of 0 or larger are retained.
- Variables with no meaning for calculations (i.e. country code, survey conduction related questions) are deleted.
- Dimension: 1500 rows, 241 columns

Median Imputation, Standardization, Data Split

- Median Imputation: Many demographic variables are categorial, the remaining variables in Model 1 and 2 are normally distributed.
- All but some demographic categorial variables are standardized.
- Data is split into training (80%) and test data (20%).

Dependent Variable: happy/not happy

- WVS: variable V10 (from 1: very happy to 4: not at all happy)
- Split V10 into "happy" for 1 and 2, and "not happy" for 3 and 4.
- The variables in Model 1 and 2 are uncorrelated with happiness.

- **Caveats**: Imbalanced positive/negative DV, external validity, social desirability, subjective survey responses, no comparison over time

3 Models

Model 1: Bhutan Happiness Index

- This model is an approximation of the Bhutan Happiness Index which originally consists of nine factors and several indicators.
- **Independent Variables:**
- Time Use (V6), Health (V11), Psychological wellbeing (V23), Living Standard (V59), Ecological diversity and resilience (V78), Cultural diversity and resilience (V79), Good governance (V115), Community vitality (V213), Education (V248)

Model 2: Demographic Model

- This model consists of nine demographic variables
- **Independent Variables:** Martial status (V57), Number of children (V58), Employment status (V229), Sector of Employment (V230), Income (V239), Gender (V240), Age (V241), Education (V248), Ethnic group (V254)

Model 3: Full Model

- Independent Variables: All 240 variables

Model 4: Empty Model

- Independent Variable: Intercept only

4 Logistic Classification, LDA, QDA and Lasso

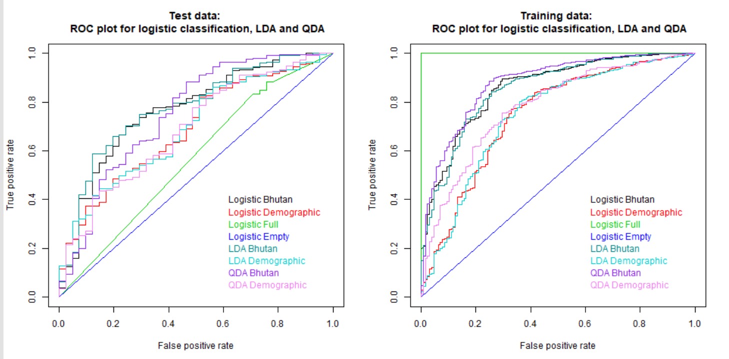

Fig. 1. ROC for test and training data for the logistic models; LDA and QDA for Bhutan and Demographic model; *Prediction accuracy:* Best Logistic: Bhutan, AUC: **0.766**, Classification Error: 0.130; Best LDA: Bhutan, AUC: **0.773**, Classification Error: 0.133; Best QDA: Bhutan, AUC: 0.753, Classification Error: **0.12.**

- Best Lasso: λ = min, 49 predictors retained, Class. error: **0.126**
- Variable importance with λ = 1 SE: Family (V4), Health (V11), Life Satisfaction (V23), Sufficient food for family in last year (V188)

5 Random Forest, Bagging and Boosting

- Bagging ($m = \sqrt{p} = 15$) using the full model returns the highest prediction accuracy for the test data with an AUC of **0.823**.
- The best boosted and tune classification tree results in a ROC of 0.8865 (Fig. 2). The prediction accuracy for the test data returns an AUC of **0.8134.**

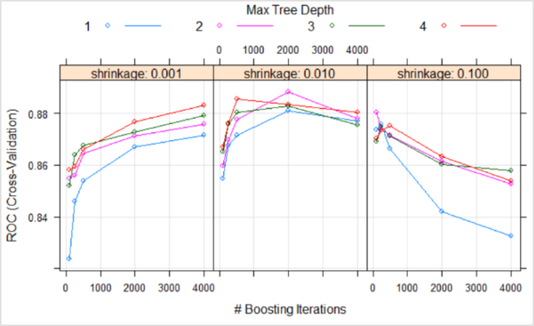

Fig. 2. Boosted and tuned classification tree to predict happiness

6 Support Vector Machine

Support Vector Machine	Linear Kernel		Polynomial Kernel		Radial Kernel	
	ROC Train	AUC Test	ROC Train	AUC Test	ROC Train	AUC Test
Bhutan	0.851	**0.7873**	0.855	0.7432	0.8245	0.7475
Demographic	0.632	0.5929	0.6925	0.7461	0.6457	0.617
Full	0.8388	0.7473	0.8385	0.6478	0.7655	0.7384

Fig. 3. Support vector machine with different kernels: Linear, polynomial and radial. *Prediction accuracy*: The linear kernel for the Bhutan model displays with 0.7873 the largest area under the curve (AUC).

7 Conclusion

Best Models: Lasso for the full model with λ = min retains 49 predictors and returns a classification error of 0.126. Bagging using the full model returns an AUC of 0.823 for the test data. Thus, using the appropriate methods, the full model returns the highest prediction accuracy.

8 References

[1] Centre for Bhutan Studies & GNH Research. (2016). A Compass Towards a Just and Harmonious Society: 2015 GNH Survey Report. Thimphu, Bhutan.; [2] Costanza et al. (2014). Time to leave GDP behind. Nature, 505, 283–285. [3] Inglehart, et al. (2014). World Value Survey: Round Six - Country-Pooled Datafile Version. Madrid. [4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An Introduction to Statistical Learning. Design (8th ed.). New York: Springer Science+Business Media.

Appendix

Logistic Classification	**Classification Error Test Data**	**Area under the curve (AUC) Test Data**
Bhutan Model	**0.13**	**0.7660797**
Demographic Model	0,.146667	0.6898955
Full Model	0.233334	0.5643658
Empty Model	0.136667	0.5

Lasso	**Error Test Data**	**Retained Predictors**
λ = 1 SE	0.126667	5
λ = min	**0.12**	49

The five variables retained in the Model with λ = 1 SE are: Intercept, Important in Life: Family (V4), State of health (V11), Satisfaction with your life (V23), In the last 12 month, how often have you or your family: Gone without enough food to eat (V188).

	Model	**ROC**	**Classification Error, Test Data**
Best logistic model	Bhutan	0.766	0.130
Best LDA	Bhutan	**0.773**	0.133
Best QDA	Bhutan	0.753	0.12
Lasso λ = min	47 retained	-	0.12
Lasso λ = 1 SE	5 retained	-	0.1266

LDA/QDA	**Classification Error (LDA) Test Data**	**Classification Error (QDA)**	**AUC LDA Test Data**	**AUC QDA Test Data**
Bhutan Model	0.133	**0.12**	**0.77286**	0.7533666
Demographic Model	0.15	0.156667	0.694039	0.6893304

	Random Forest Optimal ROC train		**Bagging ROC train**	**AUC train**	**Boosting ROC train**	
Full Model	mtry = 15	0.8463 AUC: 1	mtry = √240 = 15	0.8463 AUC: 1	mtry = 240	0.8374 AUC: 1
	Random Forest Optimal AUC test		Bagging AUC train	AUC test	Boosting AUC test	
Full Model		AUC: 0.8231		AUC: 0.8231		AUC: 0.7894

Support Vector Machine	**Linear Kernel Classification ROC Train Data**	**Error Test**	**Polynomial Kernel Classification ROC Train Data**	**Error Test**	**Radial Kernel ROC Train**	**ROC Test**
Bhutan Model	0.851	0.1366	0.855	0.12	0.8245	**0.116**
Demographic Model	0.632	0.1366	0.6925	0.1366	0.6457	0.1366
Full Model	0.8388	0.1366	0.8385	0.1333	0.7655	0.1366

Support Vector Machine	**Linear Kernel**		**Polynomial Kernel**		**Radial Kernel**	
	ROC Train	Class. Error Test	ROC Train	Class. Error Test	ROC Train	Class. Error Test
Bhutan	0.851	0.1366	0.855	0.12	0.8245	**0.116**
Demographic	0.632	0.1366	0.6925	0.1366	0.6457	0.1366
Full	0.8388	0.1366	0.8385	0.1333	0.7655	0.1366