Final Data Analysis

Due on Nov 19 at 11:59 pm EST

For this project you will take the role of a consultant working at the United Nations to explore what factors affect the happiness living in each country. They have provided you with data obtained from surveys on happiness by a third-party organization and demographic data from its own UN database. For ease of your work, the data cleaning process was done in advance, so these data were combined in a single dataset and missing data were imputed. For this project, you may analyze the data as if there were no missingness involved.

About the Project

The relevant committee in the United Nations would like to know what factors are important as to affect people's happiness, and how their effects are. They also would like to have a model that can predict people's happiness based on relevant information. It is up to you to decide what methods you want to use (frequentist or Bayesian or a combination) to answer these questions, and implement them to help identify which factors and possible interactions are at play and predict future happiness data.

In Part I, your main aim is to build a (generalized) linear model that could explain how other variables affect happiness. You should be able to interprete the results to the staff at United Nations who might not know much about statistics. In Part II, you try your best to give accurate predictions on the validation set, and you may try more complex models that are not as interpretable as linear models.

You will have three data sets: a subset for training, a subset for testing, and a third subset for validation. You will be asked to do data exploration and build your model (or models) initially using only the training data. Then, you will test your models on the testing data. The final rank is based on the performace of your models on the validation data. We are challenging you to keep your analysis experience realistic, and in a realistic scenario you would not have access to all three of these data sets at once.

All members of the team should contribute equally and may be asked to answer any questions about the analysis at the final presentation.

For your analysis create a new Rmd file or R Notebook file and update accordingly rather than editing this. Your write up should not have any of the instructions, for example. Figures should be labeled appropriately and report numbers using significant digits. This file may be updated so do not edit this document.

Code

In your write up your code should be hidden (`echo = FALSE`) so that your document is neat and easy to read. However your document should include all your code such that if I re-knit your Rmd file I should be able to obtain the results you presented. If there is any code that you wish to highlight you may included it, but it should contribute significantly to your write up that should be directed to the United Nations.

Get Started

This part is designed for you to warm up with the datasets and get a sense about what these datasets look like. You do not need to include this part in your final report.

Reading the data

To get started, read in the dataset `happiness_data.csv`. This is the dataset where you have both the predictors and response (`Happiness`). Then, read in the dataset `happiness_valid.csv`. You still have all the predictors there, but you do not have access to the response variable. Your task is to predict the response variables and submit a csv file (see `sample_submission.csv`) with two columns: country code (labeled `Code`) and predicted response (labeled `Happiness`). Note that the order of rows in this csv file might be different from that of validation set, and is sorted alphabetically according to the country code. Information about countries can be found in `CountryInfo.csv`.

You may find useful information in `Codebook.md`.

Split the training data and test data

The dataset includes happiness data from 2011 to 2018, yet not all countries have data for all 8 years. How many countries have complete data through these years? For simplicity of this report (you might not want to do this in a real-work project and all countries are important), filter out data about these countries only. Split the dataset into a training dataset and test dataset, with the training dataset including data from 2011 to 2017 and test dataset including data in 2018. Save these datasets as `happiness_train.csv` and `happiness_test.csv` for future use.

Self check: Both datasets should include 115 countries, so the training dataset should have 805 rows and the test dataset should have 115 rows.

You may assume that all countries that appear in `happiness_valid.csv` are within these 115 countries.

Overview of Workflow

Once you have created the training set and test set, all your modeling process (EDA, model fitting, model selection, etc) should only involve the training data. Use the test set for model evaluation but not for model training to avoid overfitting. When you are finally satisfied with your model, train your model on both training and test data combined, and predict happiness using the validation set. You can upload the results on Kaggle before the deadline (up to 20 times per day), and the performance on 25% of the validation set will be revealed. Finally, after the deadline, your models will be evaluated on the rest (75%) of the validation set and ranked accordingly.

The evaluation will be based on the root of mean squared error, defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Or, shown as the following code:

```r
RMSE <- function(y_true, y_pred) sqrt(mean((y_true - y_pred)^2))
```

Part I - (Generalized) Linear models

Write this part in `Part-I.rmd` and knit to `Part-I.pdf`.

The written assignment should consist of five parts:

1. Introduction: Summary of problem and objectives for this part.

2. Exploratory data analysis: Please include three or more correctly labeled graphs and an explanation that highlight the most important features that went into your model building.
3. Development and assessment of an initial model

- Initial model: What is the initial model that your group built? Does it look good?
- Model selection: Did you make modifications to your model (e.g. variable selection, add interaction, etc.)? Did you find any potention problems with your model? Repeat this process until you are satisfied with your model.
- Model diagnosis: Are there still potential problems with the model? Residual/diagnostic plots might help.
- Model summary: Include a nice-formatted table with the variables in use and provide estimates and CIs for the coefficients. Provide interprations of how the most important variables influence the (median) happiness giving a range (CI). Highlight important findings and potential limitations of your model.

4. Summary and conclusion

Provide a report (1-2 pages) to the United Nations about your findings. The United Nations appreciate all your findings, but in case you could not find one, below are some topics you might find interesting. *You do not need to follow the following topics. Some topics might not be apporpriate to your model. These are more of a hint rather than instructions.*

- Utopia, coined by Sir Thomas More in 1516, refers to a possible community or society that possesses nearly perfect qualities for its citizens. Although there is no utopia on the world, you can make up one by combining the best predictor values in the dataset into a single country. What is the predicted happiness for that utopia? Does it exceed 10? Oppositely, what is the predicted happiness for a "dystopia"? Could it be below 0?
- What are the most important factors that affect happiness? Does it vary across the years? What interactions are important?
- Although many factors contribute to happiness, some are more significant than others. It might be possible that people in different countries view happiness differently - some may value more on their health and others might value more on their wealth. Are there any regional differences on what the biggest contributions are to happiness?

5. Evaluation

Apply your model on the test set and provide point estimates as well as prediction intervals. Evaluate your model with the following criteria.

- Bias
- Maximum Deviation
- Root Mean Square Error
- Coverage

Part II - Complex Models

Write this part in `Part-II.rmd` and knit to `Part-II.pdf`.

In this part, you may go all out for constructing a best fitting model for predicting housing prices using methods that we have covered this semester. You should feel free to create any new variables (such as quadratic, interaction, or indicator variables, splines, etc) and try different methods (such as tree models or ensemble models).

The written assignment should consist of five parts:

1. Introduction: Summary of problem and objectives for this part.
2. Exploratory data analysis: Are there useful information for better prediction?
3. Discussion of preliminary model Part I: How was the performance of your linear model? Are there any refinements suggested that might improve the prediction accuracy?
4. Development of the final model: Please state clearly all models that you've attempted. Are there any hyperparameters and how did you choose them? Did you use variable selection/shrinkage and why? What is the final model you've chosen? How does it outperforms other models?
5. Assessment of the final model: Apply your model on the test set and provide point estimates as well as prediction intervals. Evaluate your model with the following criteria.

- Bias
- Maximum Deviation
- Root Mean Square Error
- Coverage (Bonus)

Final Predictions Validation

Create predictions for the validation data from your final model and upload the results to Kaggle. You may refit your final model to the combined training and test data.